Protein Folding and Structure Prediction

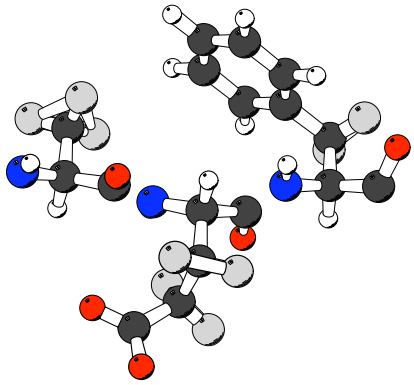
A Statistician's View

Ingo Ruczinski

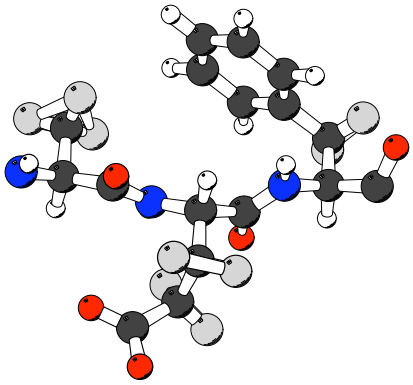
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Proteins

Amino acids without peptide bonds.

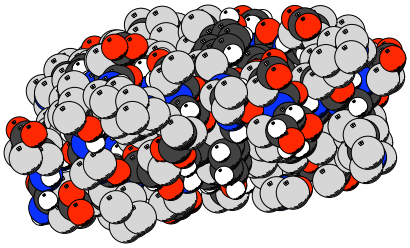


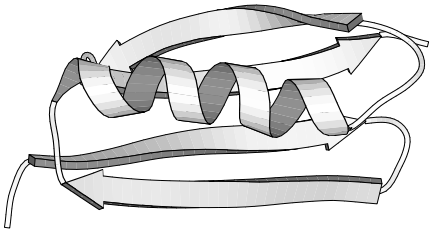
Amino acids with peptide bonds.



$\longrightarrow$ Amino acids are the building blocks of proteins.

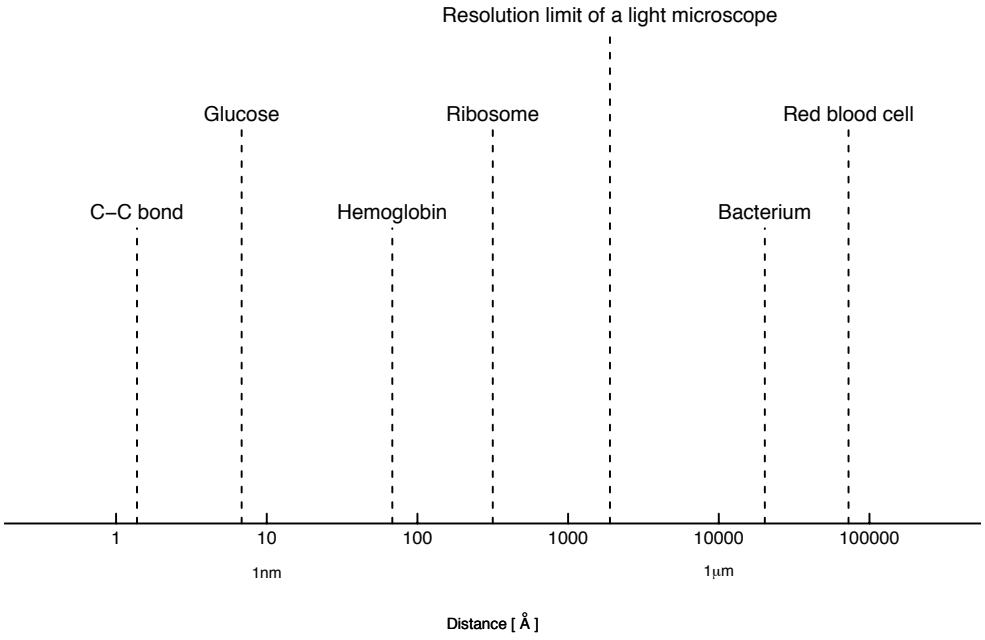
Proteins



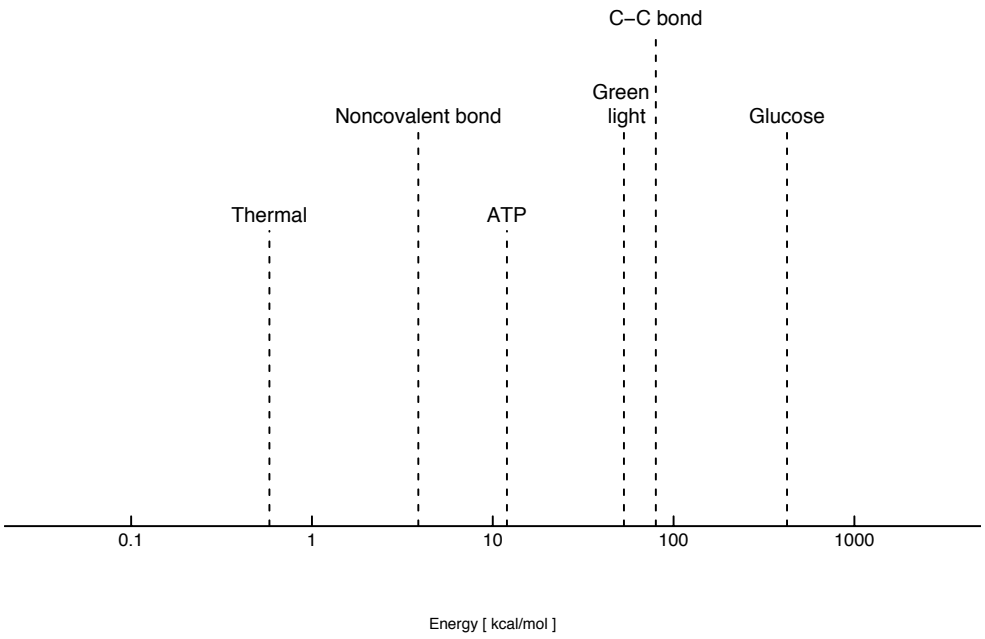


Both figures show the same protein (the bacterial protein L). The right figure also highlights the secondary structure elements.

Space



Energy



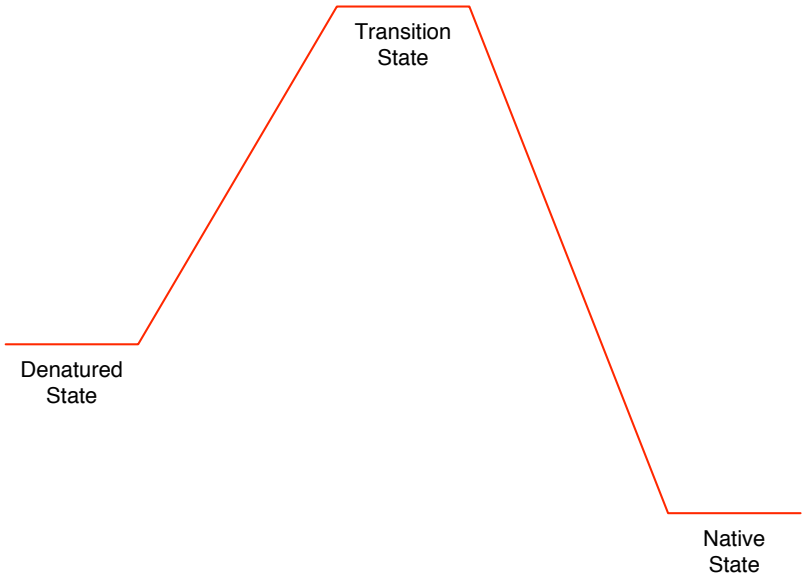
Non-Bonding Interactions

Amino acids of a protein are joined by covalent bonding interactions. The polypeptide is folded in three dimension by non-bonding interactions. These interactions, which can easily be disrupted by extreme pH, temperature, pressure, and denaturants, are:

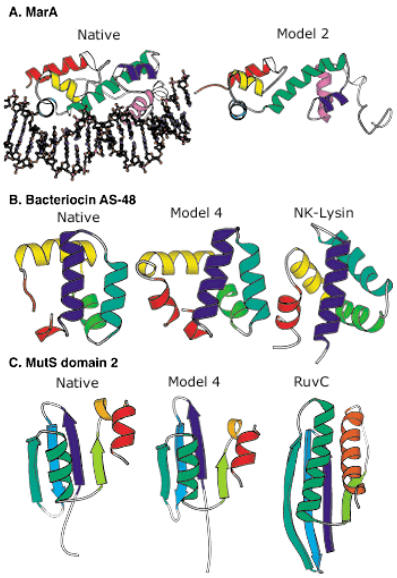
- Electrostatic Interactions (5 kcal/mol)
- Hydrogen-bond Interactions (3-7 kcal/mol)
- Van Der Waals Interactions (1 kcal/mol)
- Hydrophobic Interactions ($<$ 10 kcal/mol)

The total inter-atomic force acting between two atoms is the sum of all the forces they exert on each other.

Energy Profile

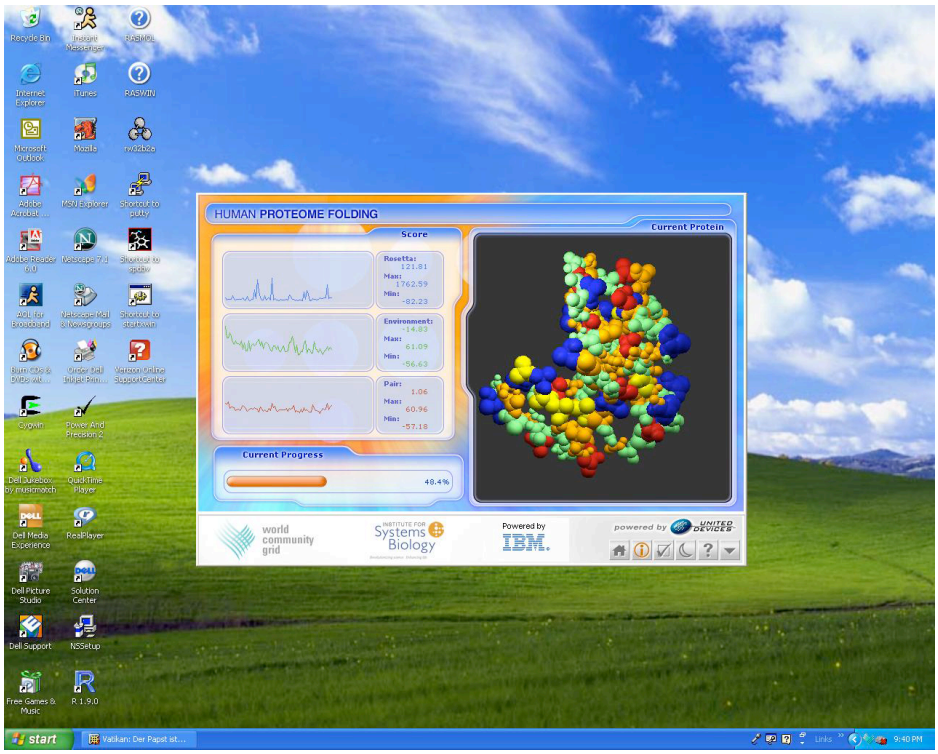


Functional Annotation

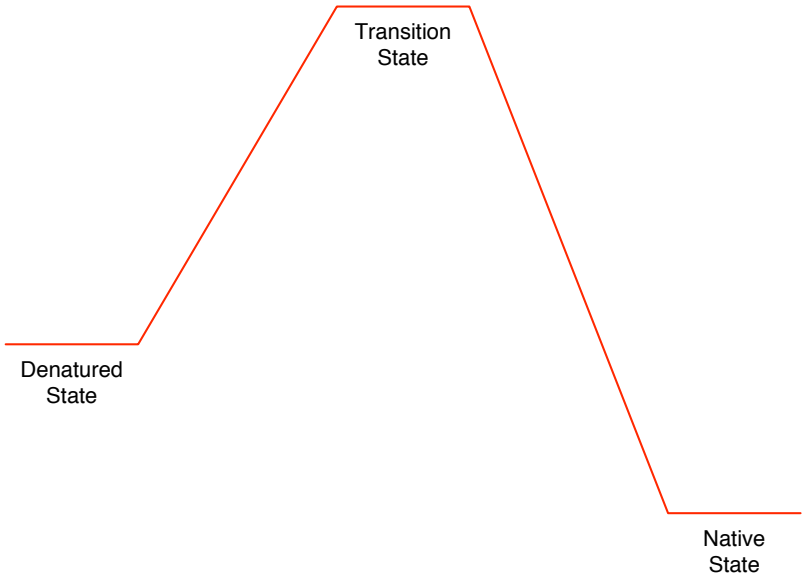


⟶ ROSETTA is used for functional annotation of genes.

www.worldcommunitygrid.org



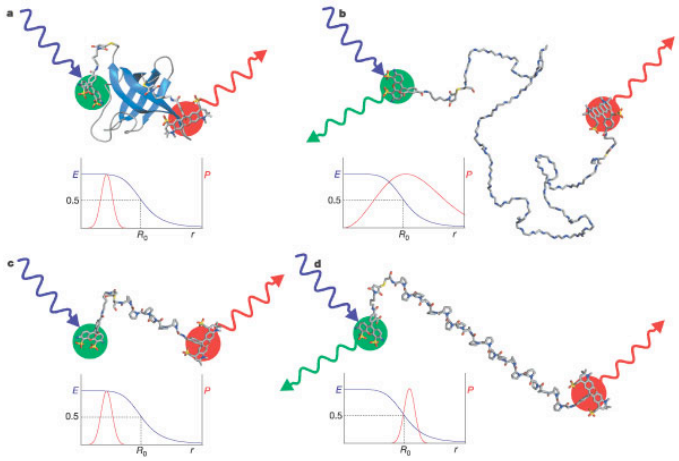
Energy Profile



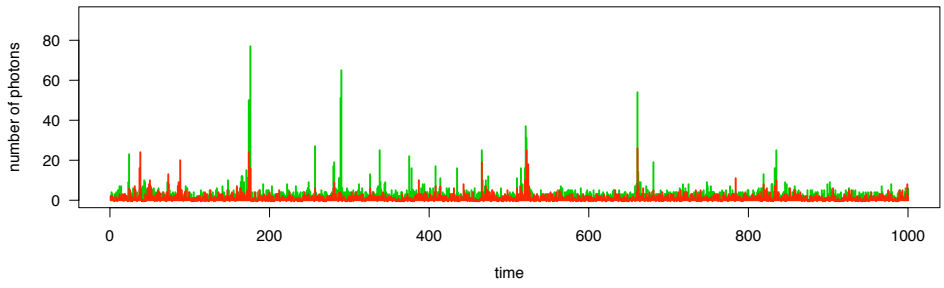
Deviations from Random Coil Behaviour

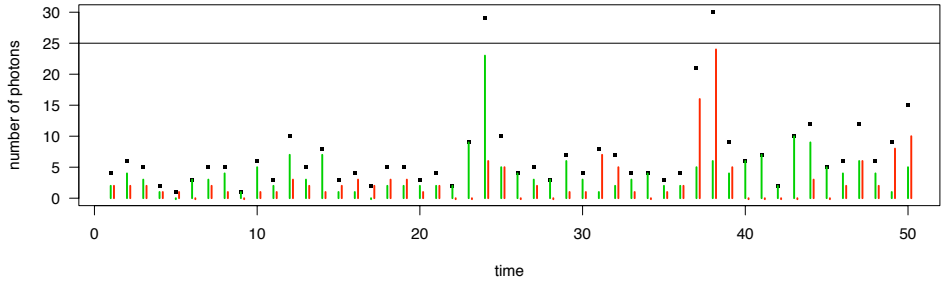
Are there site-specific deviations from random coil dimensions?

Förster Resonance Energy Transfer enables us to measure the distance between two dye molecules within a certain range. This can be used to study site-specific deviations from random coil dimensions in highly denatured peptides.

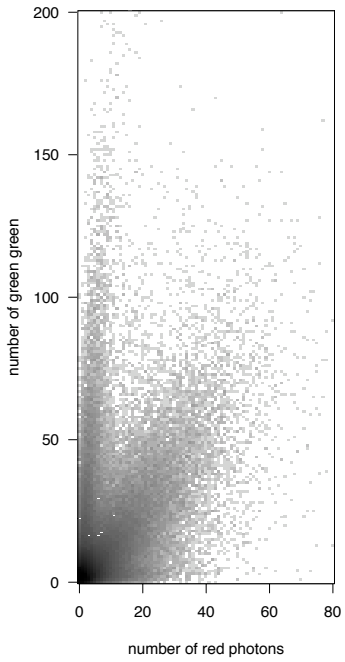


Deviations from Random Coil Behaviour





Deviations from Random Coil Behaviour

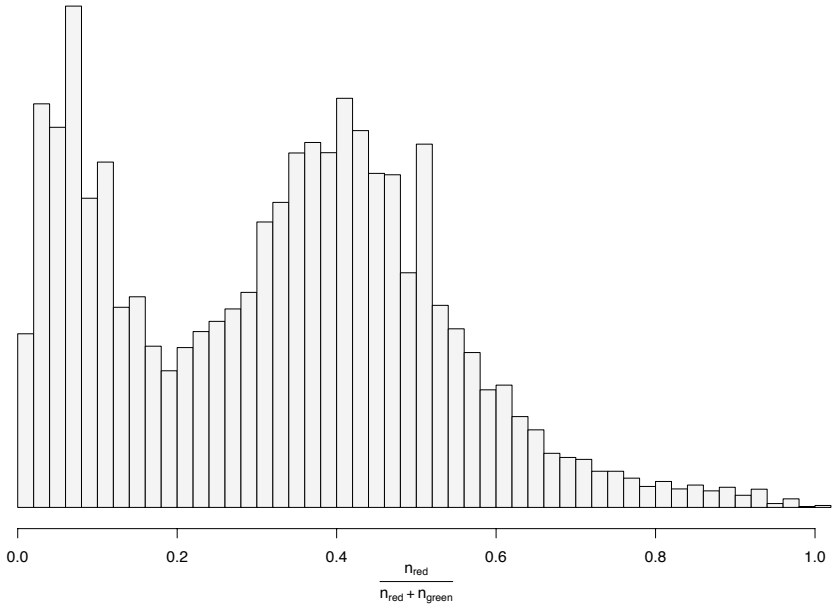


We have two underlying distributions for the green and red photons:

- One stemming from a peptide only having a donor dye.
- One stemming from a peptide being properly tagged with a donor and an acceptor dye.

Assume a photon has probability p_0 of being red in the former situation, and p_1 in the latter.

Deviations from Random Coil Behaviour



Deviations from Random Coil Behaviour

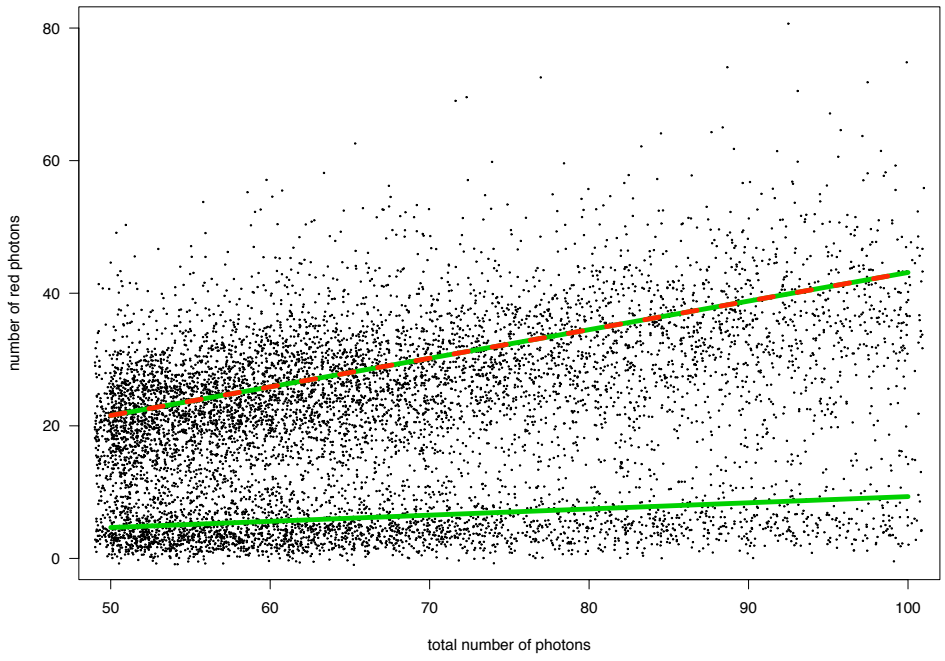
Assume we observe n_i photons at time point i. Then the number of red photons is simply Bernoulli(n_i, p_i), where p_i is either p_0 or p_1 . Assume that the probability of observing photons from a peptide without an acceptor dye at any time is p, independent of the total number of photons observed. Let X be the number of red photons. Then

$$\begin{aligned} P(X = x_i|n_i) &= P(X = x_i|n_i, p_0) \times p + P(X = x_i|n_i, p_1) \times (1 - p) \\ &= \binom{n_i}{x_i} p_0^{x_i}(1 - p_0)^{n_i - x_i} \times p + \binom{n_i}{x_i} p_1^{x_i}(1 - p_1)^{n_i - x_i} \times (1 - p), \end{aligned}$$

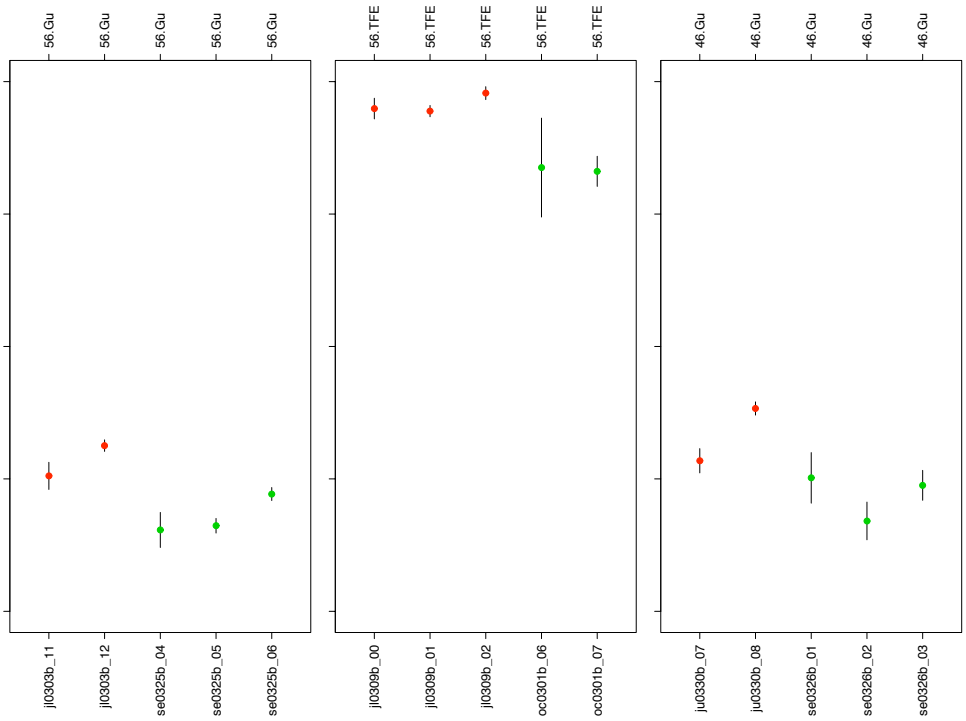
and hence

$$L(p, p_0, p_1) = \prod_{i=1}^{N} \left[\binom{n_i}{x_i} p_0^{x_i}(1 - p_0)^{n_i - x_i} \times p + \binom{n_i}{x_i} p_1^{x_i}(1 - p_1)^{n_i - x_i} \times (1 - p) \right].$$

Deviations from Random Coil Behaviour

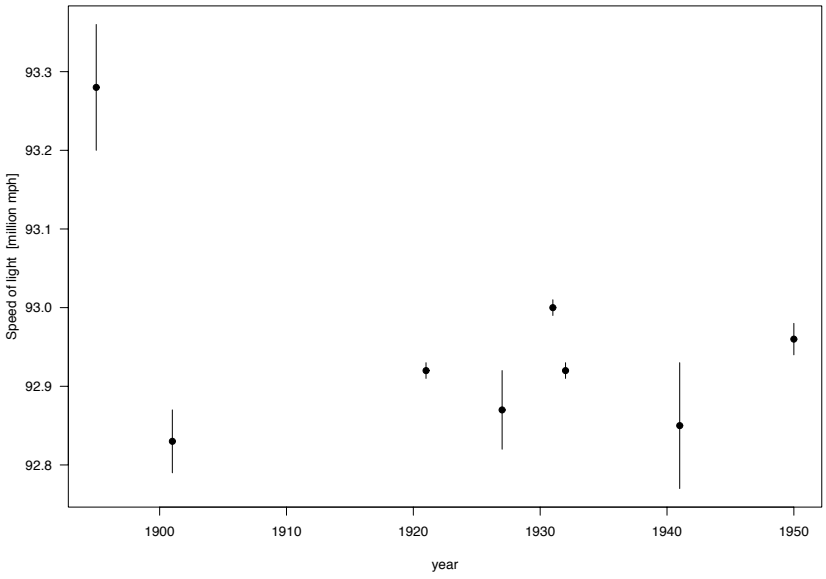


Variance Components ?



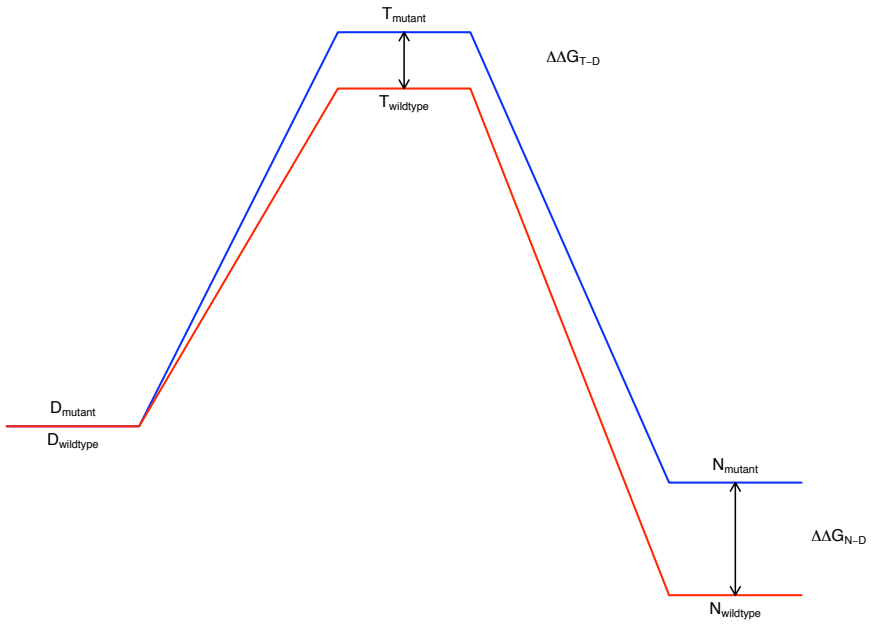
Dang...

Estimates of the speed of light, with “confidence intervals” (1895 - 1950).



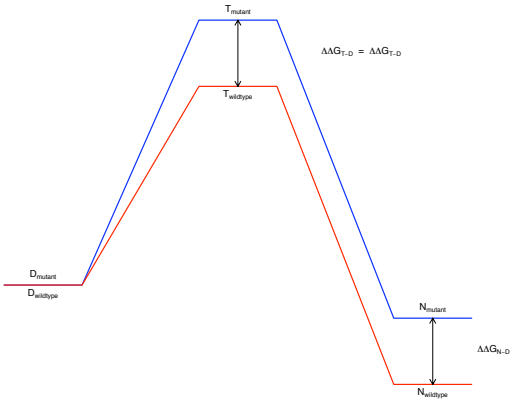
Youden (Technometrics, 1972).

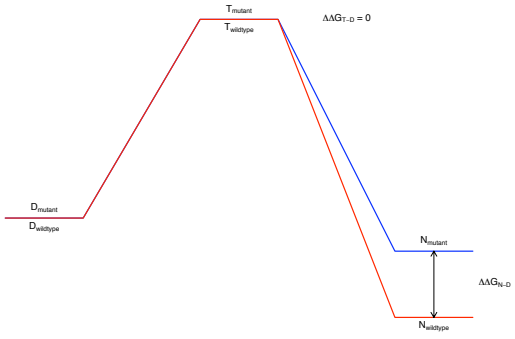
Energy Profile



$\longrightarrow$ The Φ-value is defined as the ratio $\Delta\Delta G_{T\text{-}D}/\Delta\Delta G_{N\text{-}D}$.

Energy Profile

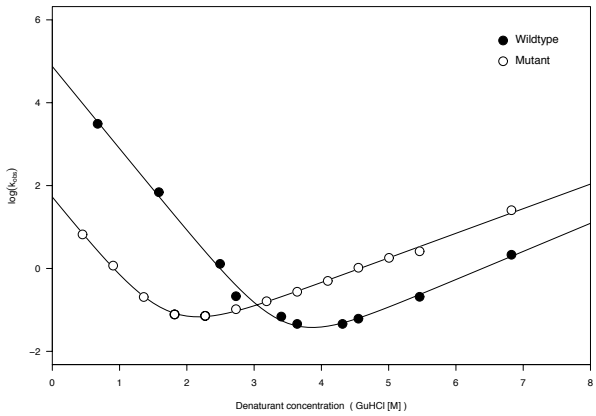




- If the part of the protein that contains the mutant amino acid is fully structured in the transition state, we have $\Delta\Delta G_{T\text{-}D} \approx \Delta\Delta G_{N\text{-}D}$, and hence $\Phi \approx 1$.
- If the part of the protein that contains the mutant amino acid is equal in denatured and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

At least this is the idea . . .

Phi-Value Estimation

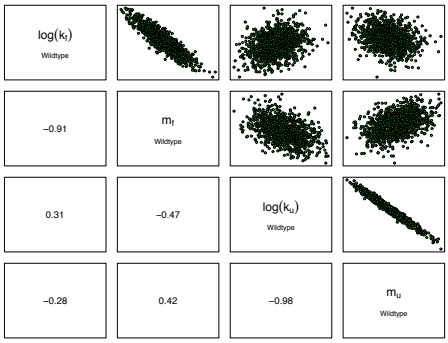


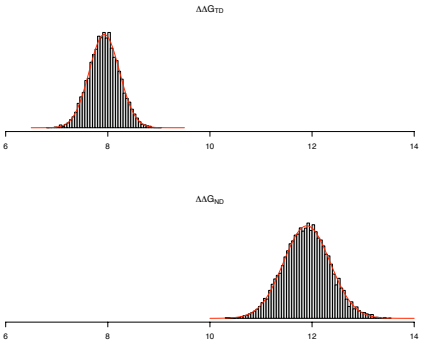
$$\log(k_{obs}) = \log\left(\exp\left[\log(k_f) + m_f \times \frac{C_{GuHCl}}{RT}\right] + \exp\left[\log(k_u) + m_u \times \frac{C_{GuHCl}}{RT}\right] \right)$$

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

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

Confidence Intervals





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

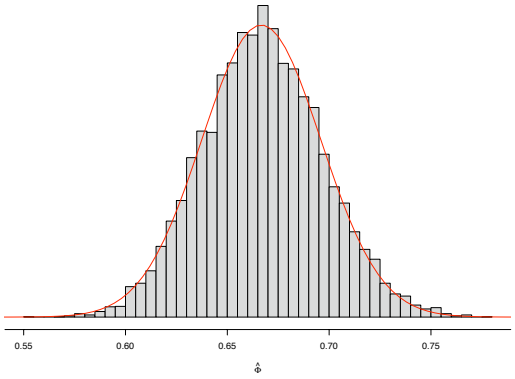
$$\begin{aligned} \sigma_1^2 &= \sigma_{F_W}^2 + \sigma_{F_M}^2 \\ \sigma_2^2 &= \sigma_{F_W}^2 + \sigma_{F_M}^2 + \sigma_{U_W}^2 + \sigma_{U_M}^2 - 2\rho_W \sigma_{F_W} \sigma_{U_W} - 2\rho_M \sigma_{F_M} \sigma_{U_M} \\ \sigma_3^2 &= \sigma_{F_W}^2 + \sigma_{F_M}^2 - \rho_W \sigma_{F_W} \sigma_{U_W} - \rho_M \sigma_{F_M} \sigma_{U_M} \end{aligned}$$

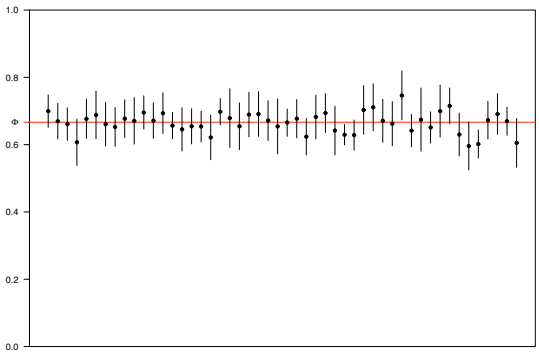
For sufficiently large $\Delta\Delta G_{N\text{-}D}$, some more math shows that the estimate for Φ is approximately normal (there is some slight abuse of the "delta method" involved).

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

Confidence Intervals

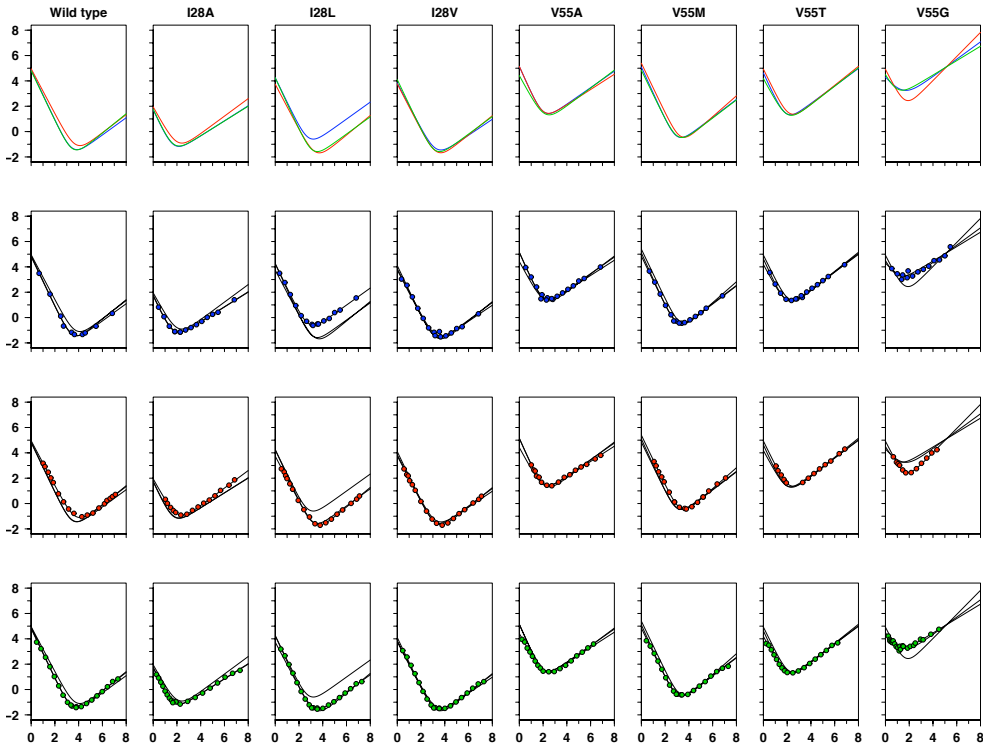
Confidence intervals for the Φ-value: $I = \left[\hat{\Phi} - t^{0.975}_{n_1+n_2-10} \times \sqrt{B}\ ;\ \hat{\Phi} + t^{0.975}_{n_1+n_2-10} \times \sqrt{B}\right]$



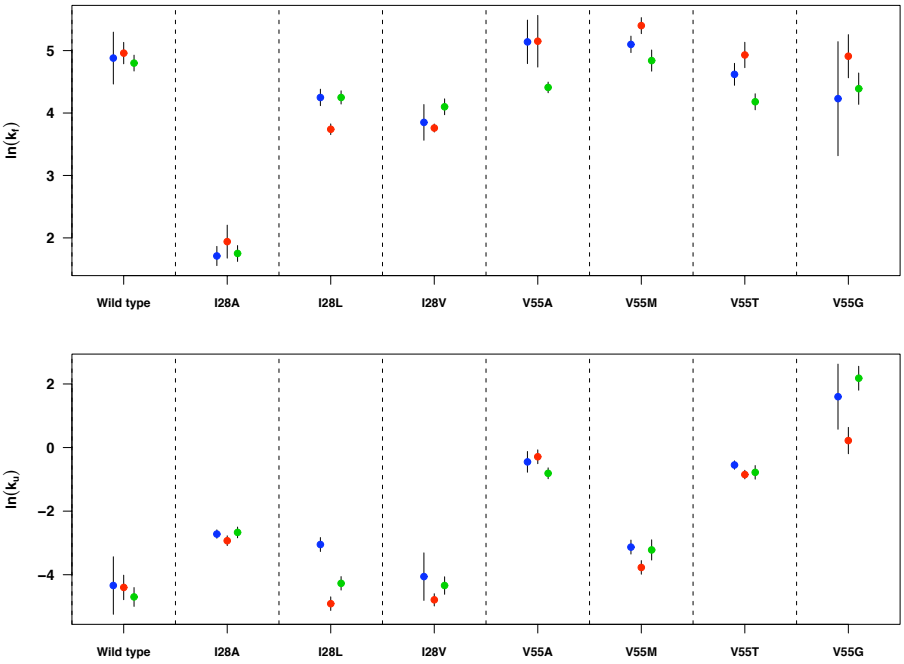


$\longrightarrow$ It is not a priori clear what the degrees of freedom in the t-quantile should be. Adding the number of data points used to fit the two chevron curves (n_1 and n_2) and subtracting the number of parameters estimated in the fitting procedure (a total of 10) however gave 95% coverage for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

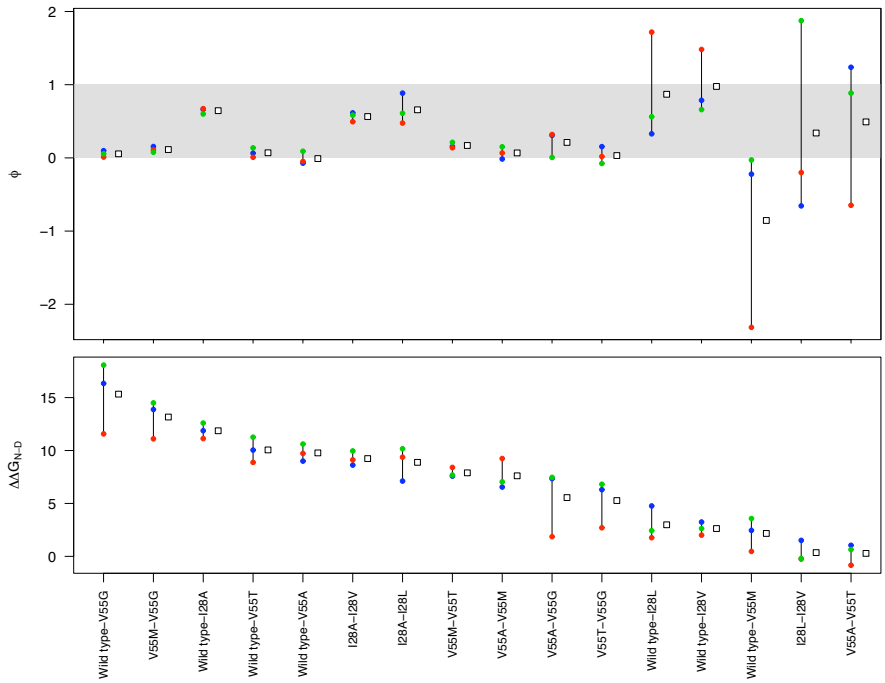
More Chevron Plots



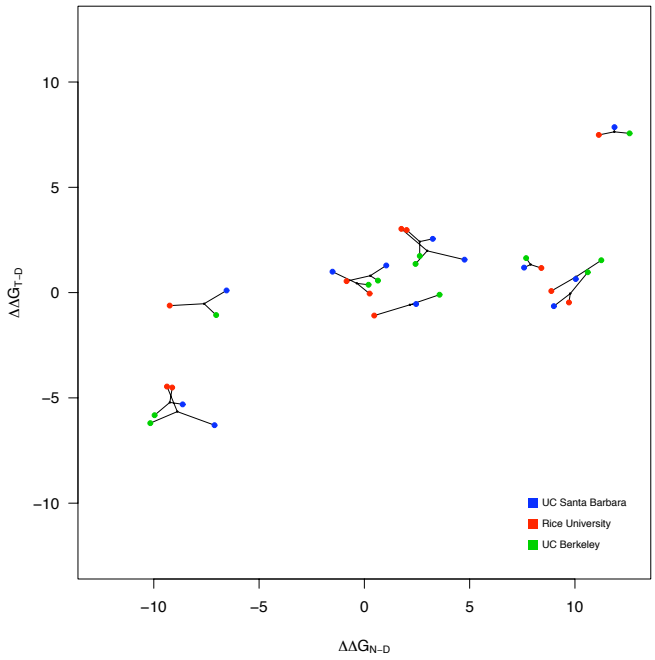
Variability



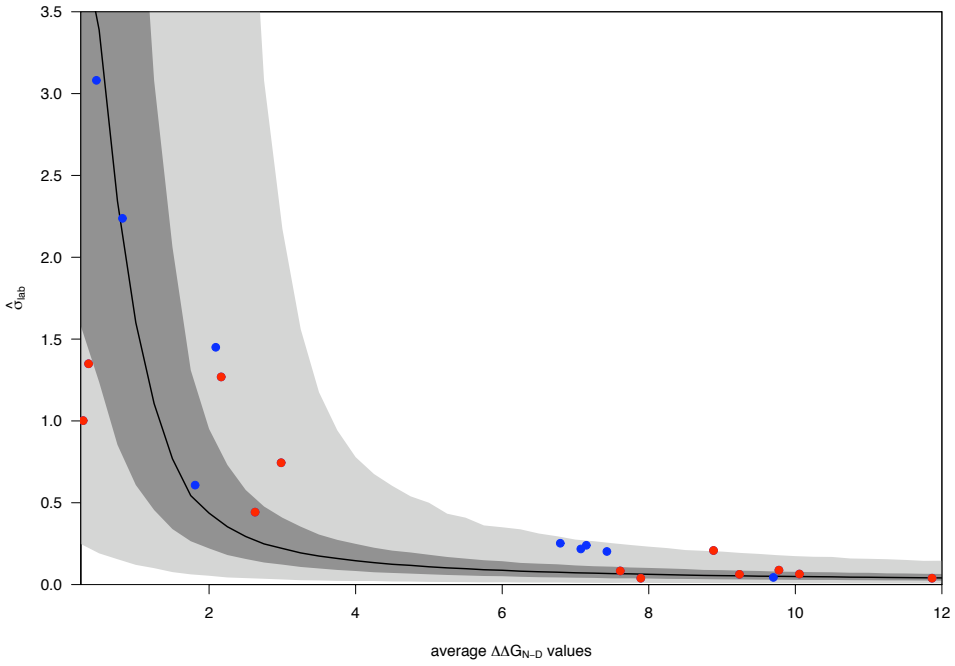
Variability



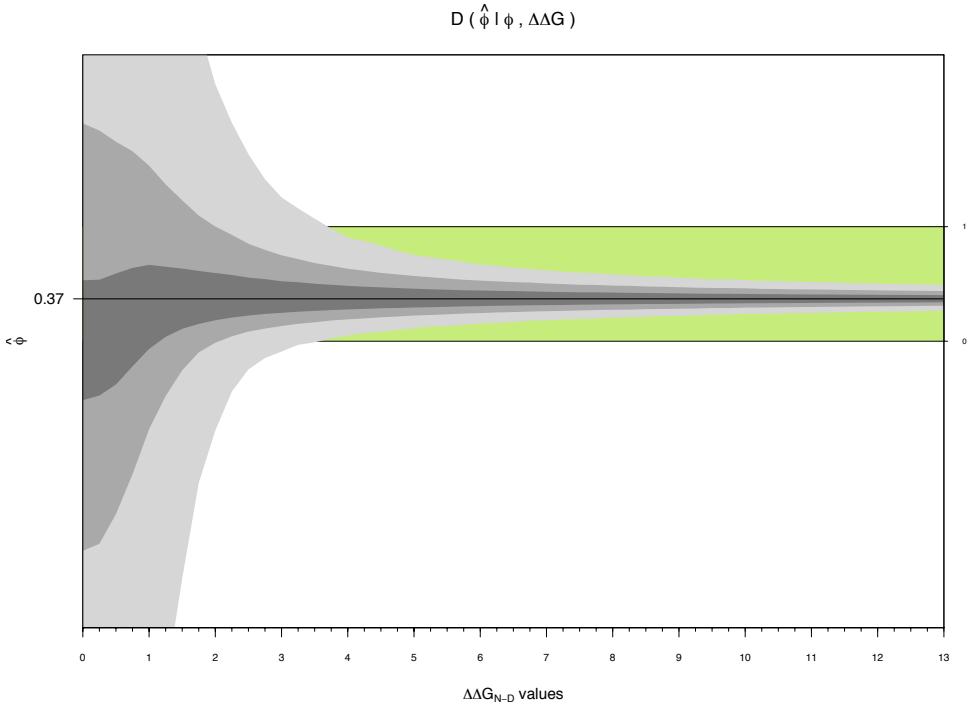
Variability



Some Simulation



Some More Simulations



Some More Simulations