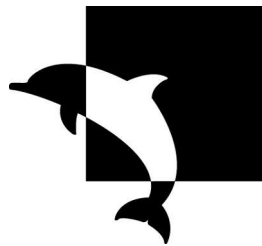


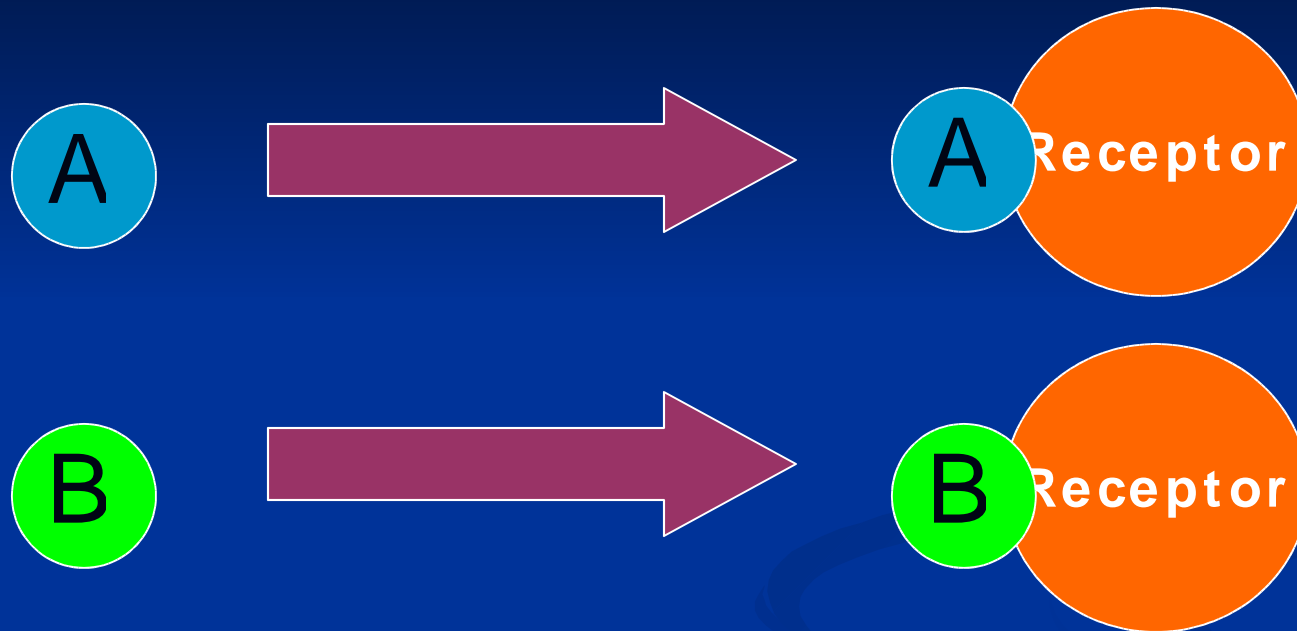
Introduction to Free Energy Calculations

Nov. 2005
Julien Michel



**University
of Southampton**

Predicting Binding Free Energies

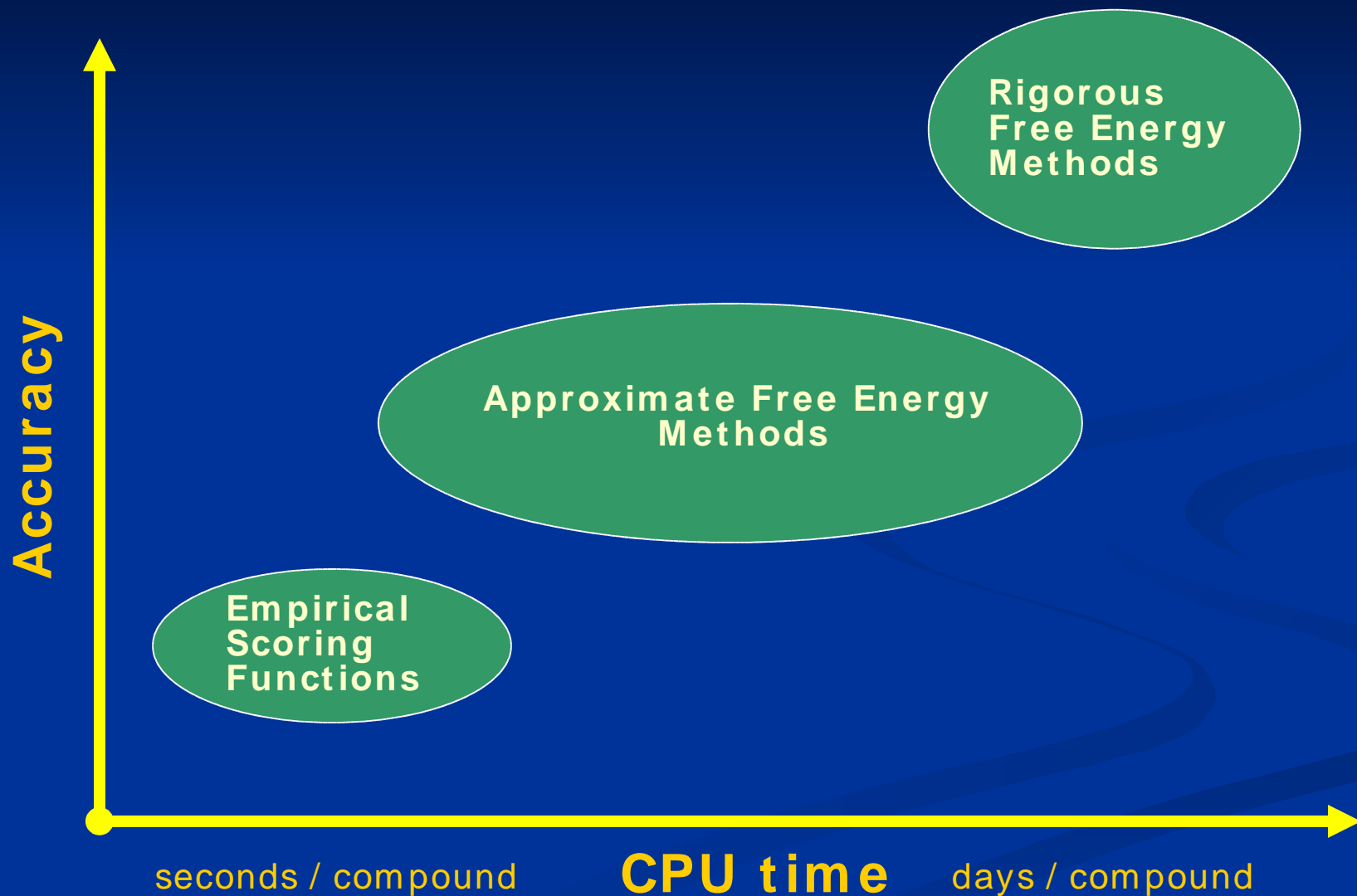


Can we predict if A or B will bind ?

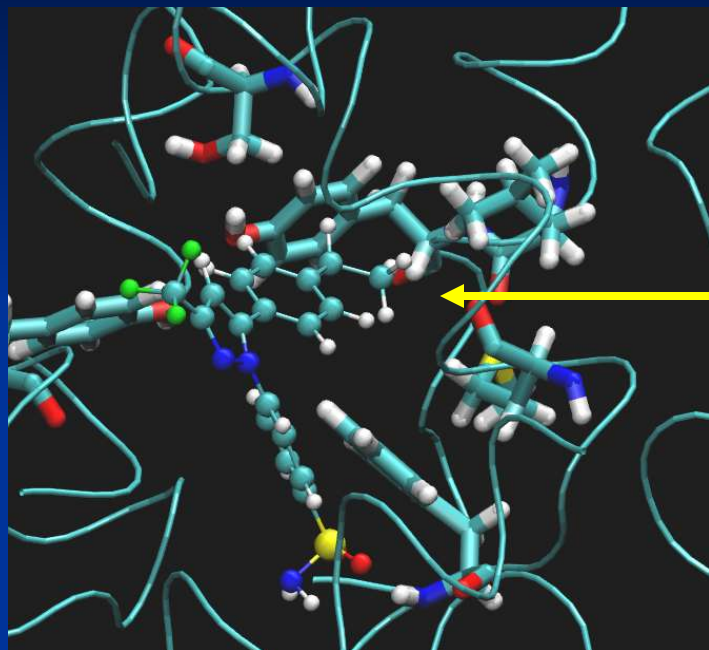
Can we predict the stronger binder ?

Can we do this reliably and quickly ?

What is in the toolbox ?



The importance of the entropy

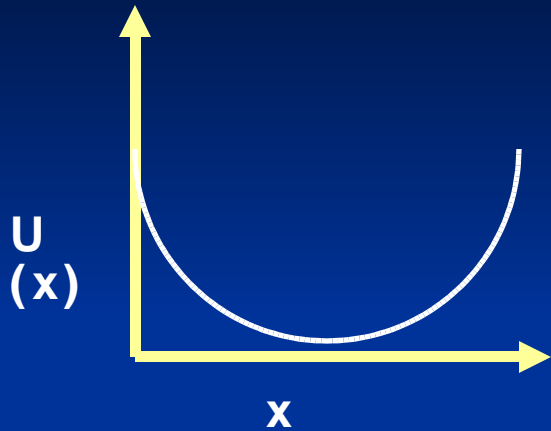


- Better Lennard Jones interactions with the protein
- But configurationally restricted
- Experimental Binding Free Energy difference : + 1.8 kcal.mol⁻¹

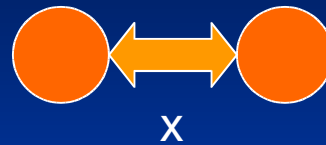
Free Energy calculations capture aspects of ligand binding that are overlooked (or crudely considered) by empirical scoring functions :

- Entropic effects
- Solvation

The concept of Phase space

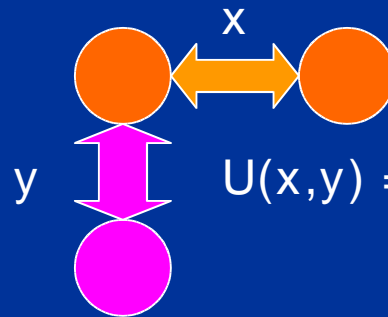


1 dimensional harmonic oscillator



$$U(x) = \frac{k_1}{2} (x - x_0)^2$$

2 dimensional harmonic oscillator



$$U(x,y) = \frac{k_1}{2} (x - x_0)^2 + \frac{k_2}{2} (y - y_0)^2$$

- Serious simulations typically requires thousands of degrees of freedom and hence the phase space is in thousands of dimensions

Statistical Physics in one equation

$$\langle P_{ens} \rangle = \frac{\int P(r^N) \exp(-\beta U(r^N)) dr^N}{\int \exp(-\beta U(r^N)) dr^N}$$

$$\langle P_{ens} \rangle = \int P(r^N) \pi(r^N) dr^N$$

Boltzmann distribution

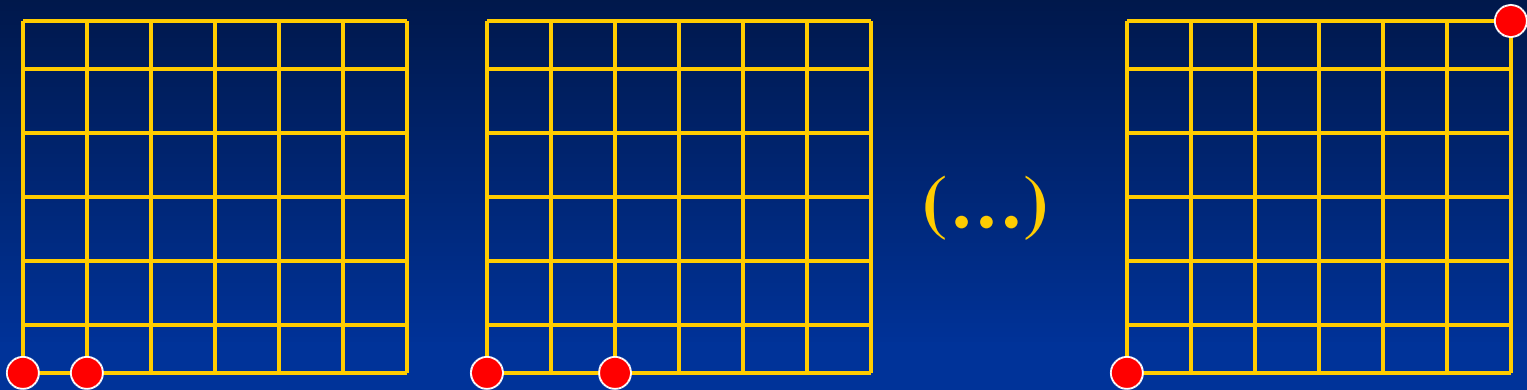
Point in Phase space

Ensemble Average of P

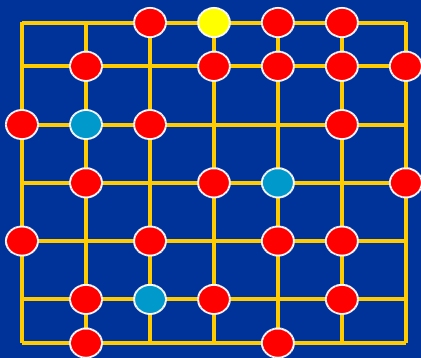
Value of P for one configuration r^N

$$P_{macroscopic} = \langle P_{ens} \rangle \quad (\text{Postulate})$$

Numerical integration : Quadrature



- Two atoms on a 7*7 checkerboard :
 $49 \times 49 = 2401$ different states contributes to eq (2)



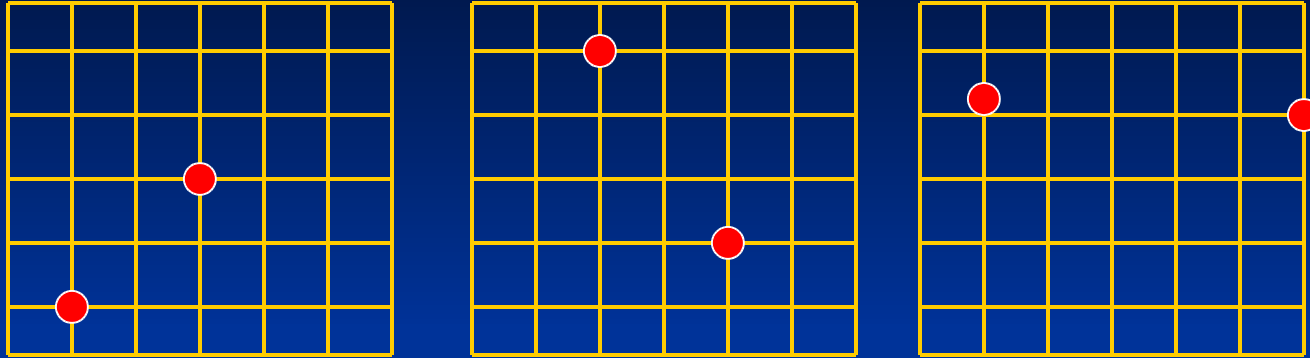
- Lot's of states !
- In eq (2), the integral is in $3N$ dimensions where N is the number of atoms simulated. And there are thousands of atoms to simulate...



Direct Integration is not feasible

Important remark : Many, many of the possible states involve overlaps of atoms and hence are highly unlikely

Numerical integration : Monte Carlo

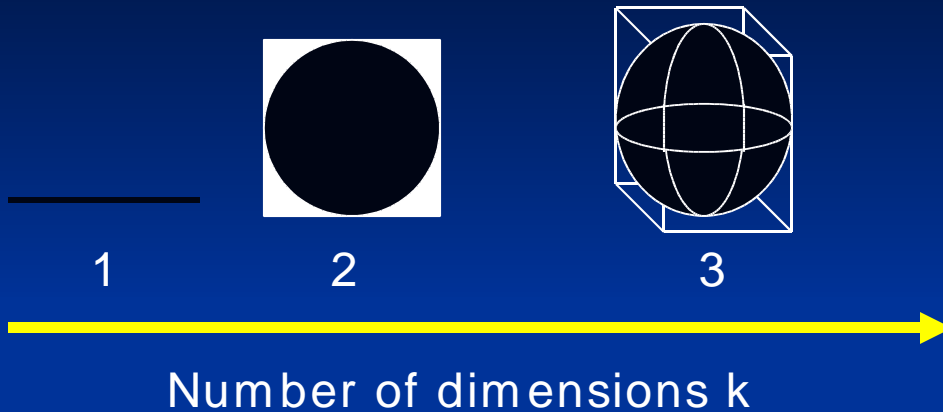


- Generate **randomly** the position of the atoms on the checkerboard, calculate the property of interest $I_s(X_i)$
- Repeat N times
- The quantity I_{est} is an estimate of the integrant /

$$I_{est} = \frac{1}{N} \sum_{i=1}^N I_S(X_i)$$

The essential advantage over quadrature techniques is that the method can be applied to systems of high dimensionality without having to build a “mesh” that covers space

What is the volume of a "sphere" ?



$$\frac{V}{V_R} = \frac{\pi^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2} + 1\right) 2^k}$$

| k | V/V_R |
|-----|----------|
| 1 | 1.00E+00 |
| 2 | 7.85E-01 |
| 3 | 5.24E-01 |
| 5 | 1.64E-01 |
| 10 | 2.49E-03 |
| 50 | 1.54E-27 |
| 100 | 1.87E-69 |

As k increases, the vast majority of the points in the k -dimension space lies outside of the sphere

Random selection of points doesn't work

Strong analogy with systems in the condensed phase. There are **few low energy states** that contributes meaningfully to the integral and **many high energy states** (e.g atomic overlaps) that do not contribute.

Importance Sampling

- Instead of drawing random points from an uniform distribution, draw points from a distribution π . The Monte Carlo integration equation becomes

$$I_{est} = \frac{1}{N} \sum_{i=1}^N \frac{I(X_i)}{\pi(X_i)}$$

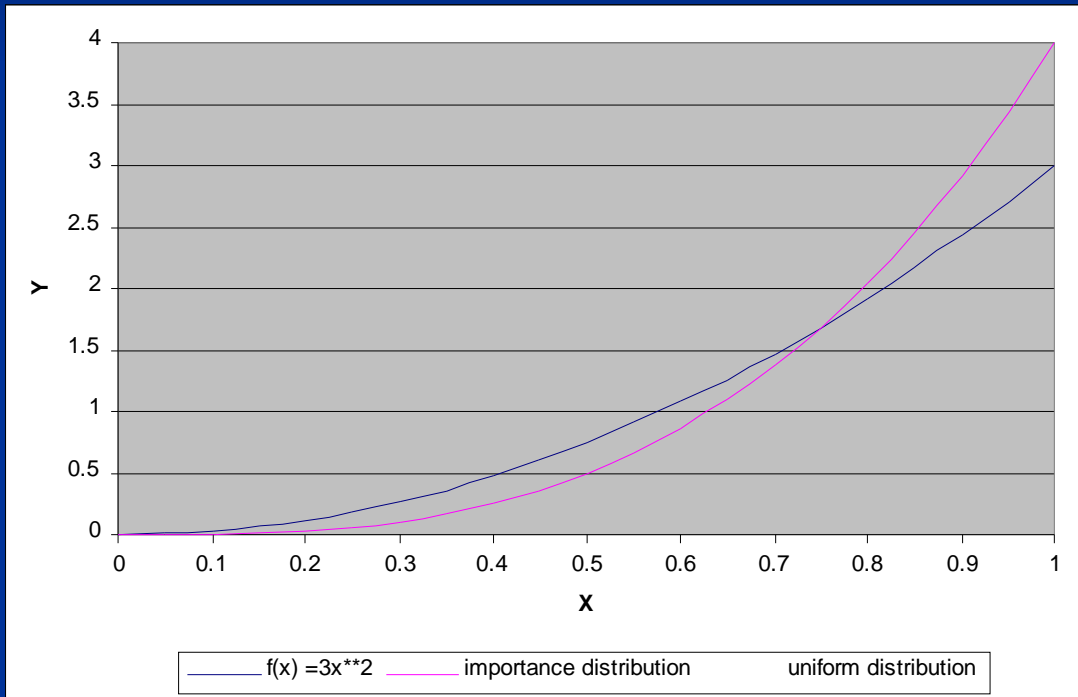
- π is selected so that points are in the region of space which contributes the most to the integrand (e.g, in the sphere)
- The bias on the selection of X_i is removed when the contribution to the integrand is estimated.

Importance Sampling : Example

$$f(x) = 3x^2$$
$$I = \int_0^1 \frac{f(x)}{\pi_k(x)} \pi_k(x) dx$$

$$\pi_0 = 1$$

$$\pi_1 = 4x^3$$



Estimate of I after drawing 100 samples with two different importance sampler

| Function | Average | Std. Dev |
|----------|---------|----------|
| π_0 | 1.027 | 0.111 |
| π_1 | 0.999 | .036 |

Importance Sampling in Statistical Physics

$$\langle P_{ens} \rangle = \frac{\int P(r^N) \exp(-\beta U(r^N)) dr^N}{\int \exp(-\beta U(r^N)) dr^N}$$

If $P(r^N)$ does not dominate the product in the numerator, then an ideal importance sampling function to estimate equation (2) is :

$$\pi(r_i) = \frac{\exp(-\beta U(r_i))}{\int \exp(-\beta U(r^N)) dr^N}$$

Problem : Impossible to draw samples from $\pi(r_i)$ without knowing the denominator, which we can't (as it involves solving directly a very difficult integral)

Markov Chains (in a nutshell)

- A Markov Chain is a set of **probabilistic rules** which governs **transitions between states** and is often represented as a **transition matrix Π**

$$\Pi = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix}$$

← Probability of moving from state 1 to 3

- Assuming Π obeys a number of mathematical properties, then the following is true

$$\pi = \lim_{n \rightarrow \infty} p^{(1)} \Pi^{(n)}$$

$p^{(1)}$ represent an arbitrary initial distribution (e.g, a randoms starting point) and $\Pi^{(n)}$ represents n applications of the transition matrix Π

This equations means : Repeated applications of the transition matrix Π converges any initial arbitrary distribution $p^{(1)}$ towards a **unique limiting distribution π**

If the transition matrix is properly constructed, we can draw samples from p without having to specify π a priori (e.g, not have to solve the bad integral)

Markov Chains Monte Carlo : Metropolis Monte Carlo (1953)

1. Start in state i
2. Attempt a move to state j with probability p_{ij}
3. Accept this move with probability

$$\alpha_{ij} = \frac{\pi_j}{\pi_i} = \frac{\exp(-\beta U_j) / Z_N}{\exp(-\beta U_i) / Z_N} = \exp(-\beta(U_j - U_i))$$

6. If the move is accepted, set $i = j$, otherwise $i = i$
7. Accumulate any property of interest $A(i)$
8. Return to 1 or terminate after N iterations

Perhaps the most important 20th century contribution to numerical science

$$A = \int \exp(+\beta U(r^N)) \pi(r^N) dr^N$$

Absolute free energy calculation :

$$\langle P_{ens} \rangle = \int P(r^N) \pi(r^N) dr^N$$

$$A = \int \exp(+\beta U(r^N)) \pi(r^N) dr^N$$

$$A = \langle \exp(+\beta U(r^N)) \rangle$$

The first term in the integral is significant compared to the second term. Regions of space that contributes to the integral are not always those that have a significant value of $\pi(r)$ **Very slow** convergence of the free energy

Anyway...What is the free energy of a cactus ?

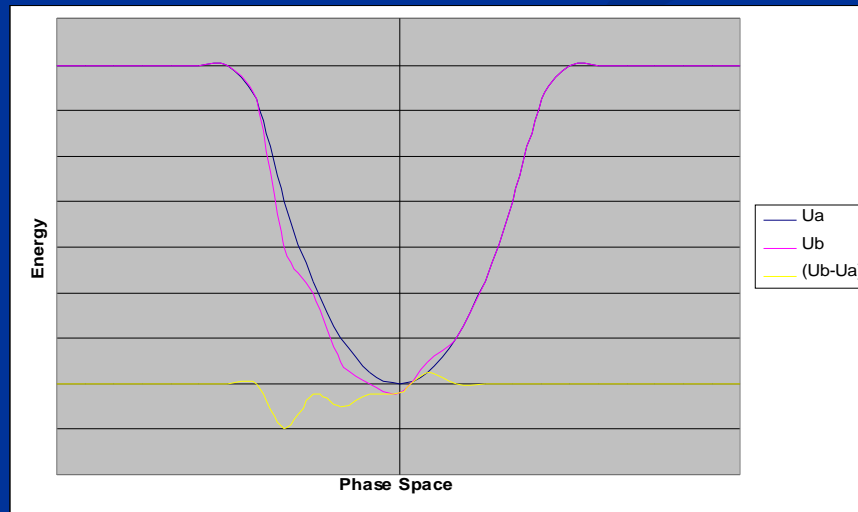
Relative free energy calculation :

Zwanzig equation (1954)

$$\Delta A_{A \rightarrow B} = -\frac{1}{\beta} \ln \left\langle \exp \left(-\beta (U_B(r^N) - U_A(r^N)) \right) \right\rangle_A$$

- If A and B are similar, then most configurations of the system A will have a similar energy to the system B and the energy difference is 0

Much better convergence because only a small number of configurations have to be generated (those that differs in energy* between A and B)



* probability would be more accurate

The coupling parameter λ :

- A and B have to be very very similar for the previous equation to yield converged results in a reasonable time

Solution, multi stage the calculation with as many intermediates potential as it takes

$$A_B - A_A = \Delta A = \sum_{k=0}^1 -\frac{1}{\beta} \ln \left\langle \exp \left(-\beta (U_{\lambda_{k+1}}(r^N) - U_{\lambda_k}(r^N)) \right) \right\rangle_{\lambda_k}$$

- The intermediates potentials U_k can be anything you want. Typically taken by linear combination of the potential for system A and B

$$U_{\lambda_k} = (1 - \lambda_k) U_A + \lambda_k U_B$$

Thermodynamic Integration :

$$\Delta A_{A \rightarrow B} = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{\partial U}{\partial \lambda} \right\rangle_{\lambda} d\lambda$$

- Another way to get the free energy difference between A and B

The quantity in brackets is the ensemble average of the gradient of the energy with respect to the coupling parameter

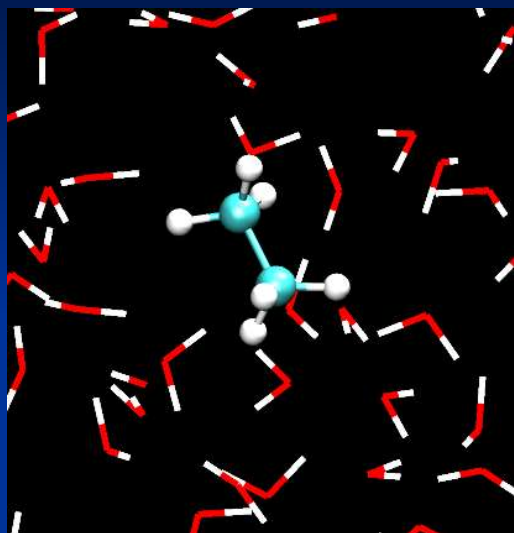
Calculated by implementing λ derivatives in the simulation code

Or by finite difference approximation

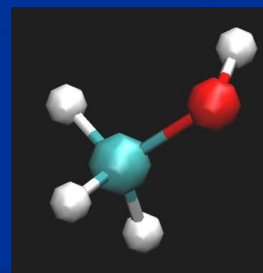
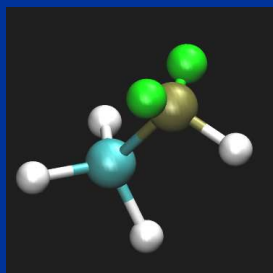
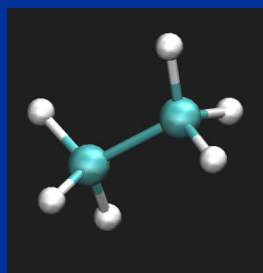
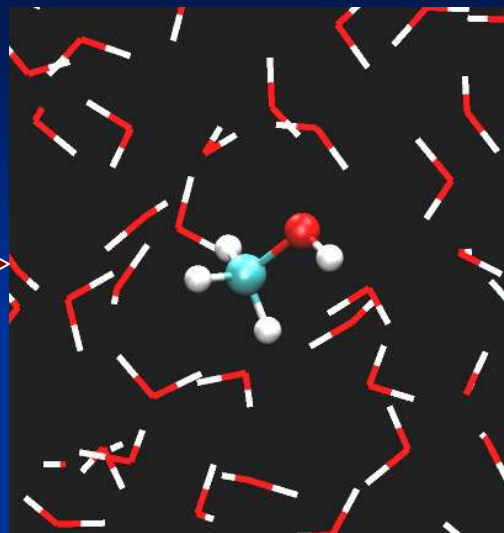
$$\frac{\partial U}{\partial \lambda} \approx \frac{\Delta U}{\Delta \lambda}$$

- If the change A to B is “big”, the gradients varies a lot and many points are necessary to obtain a good estimate of ΔA by numerical integration

Summary



ΔG ?



0.0

λ

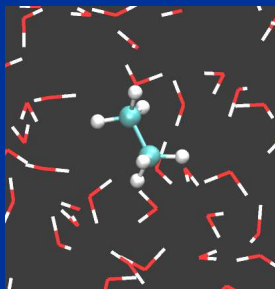
1.0



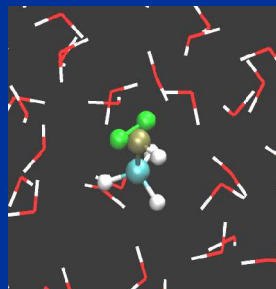
Summary

$$\Delta G = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_{\lambda} d\lambda$$

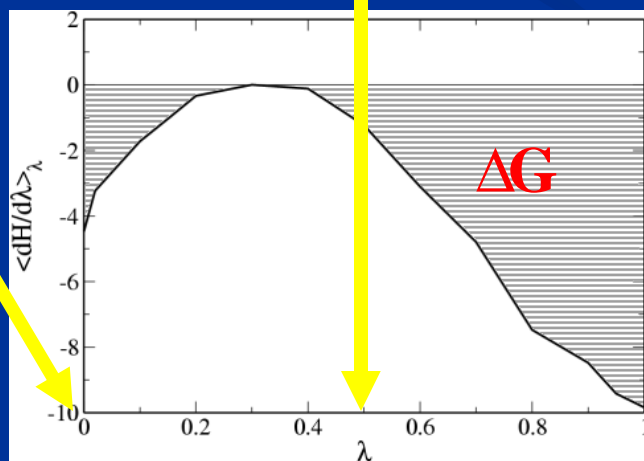
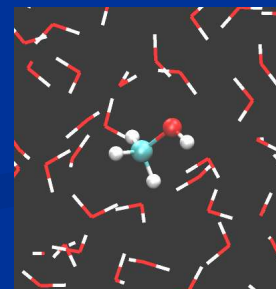
$\lambda = 0.0$



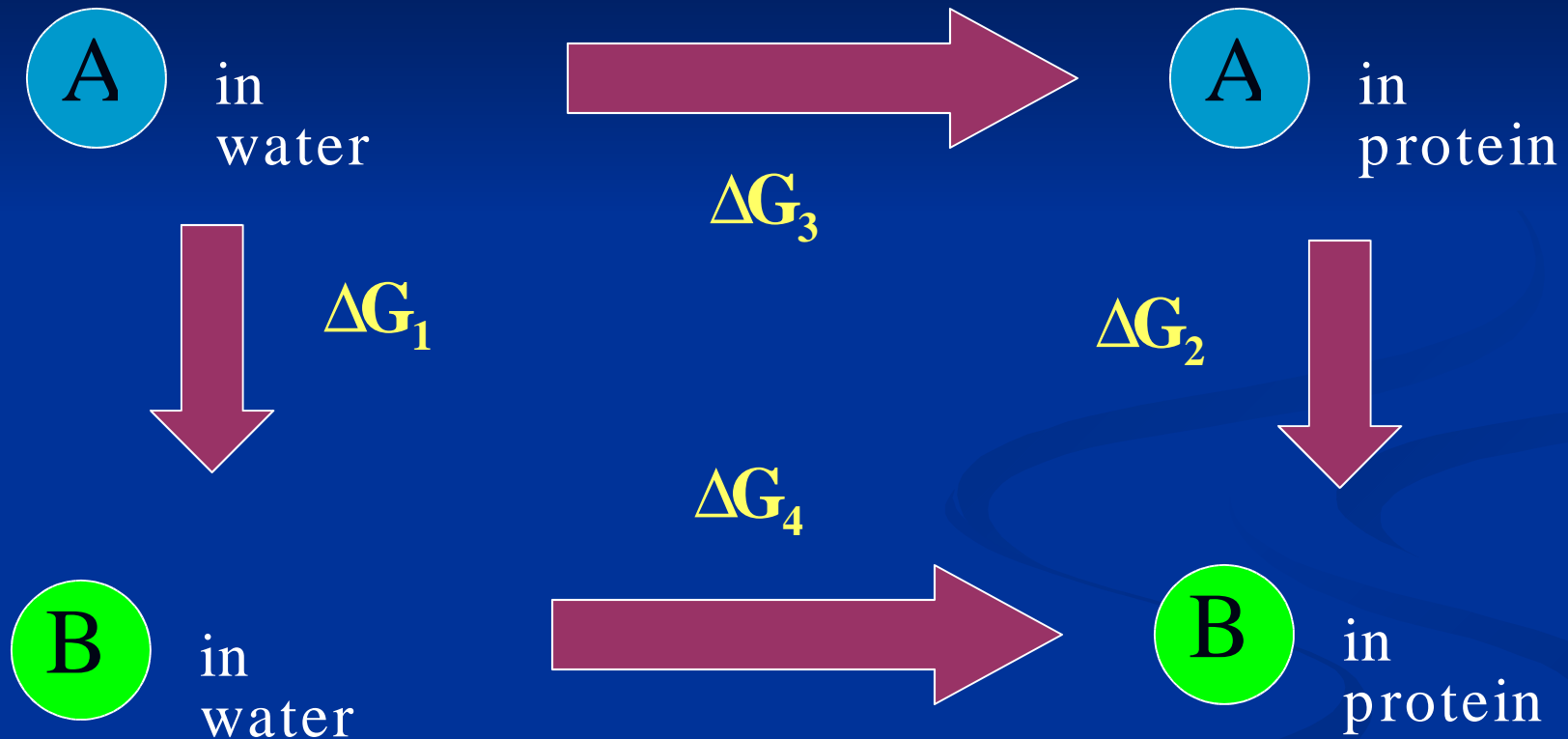
$\lambda = 0.5$



$\lambda = 1.0$



Relative Binding Free Energy Calculations by a Thermodynamic cycle



$$\Delta\Delta G = \Delta G_4 - \Delta G_3 = \Delta G_2 - \Delta G_1$$

Problems and pitfalls :

- Forcefield

The potentials U_A and U_B are approximate and this affect the quality of the results (ligand polarisation in water and binding site ?)

- Convergence

- Long simulation times required to get a good estimate of the free energy change (= many hours of CPUs)
- Many intermediate potentials have to be run to get a good estimate of the free energy change (= many CPUs)

It is impossible to know if your results have converged ! Some calculations give the same answer if you start from a different point in phase space and run with a different random number. Other never gives the same answer...This restrict the applicability of the method to system that are very similar.

Diagnosing convergence:

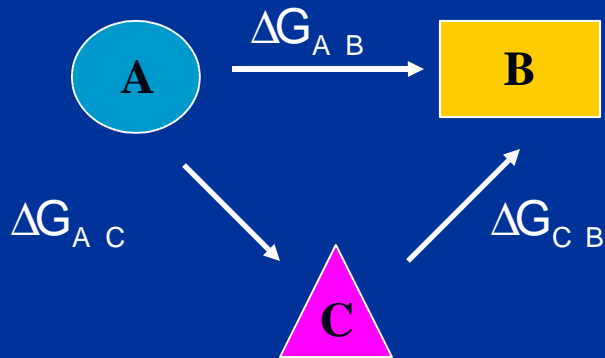
- Block averaging

Cut the raw data into chunks and analyse independently each chunk to get a free energy. Measure the variance of the resulting distribution of free energies.

- Independent runs

Run the complete simulation many times, from different starting points

- Thermodynamic cycle closure

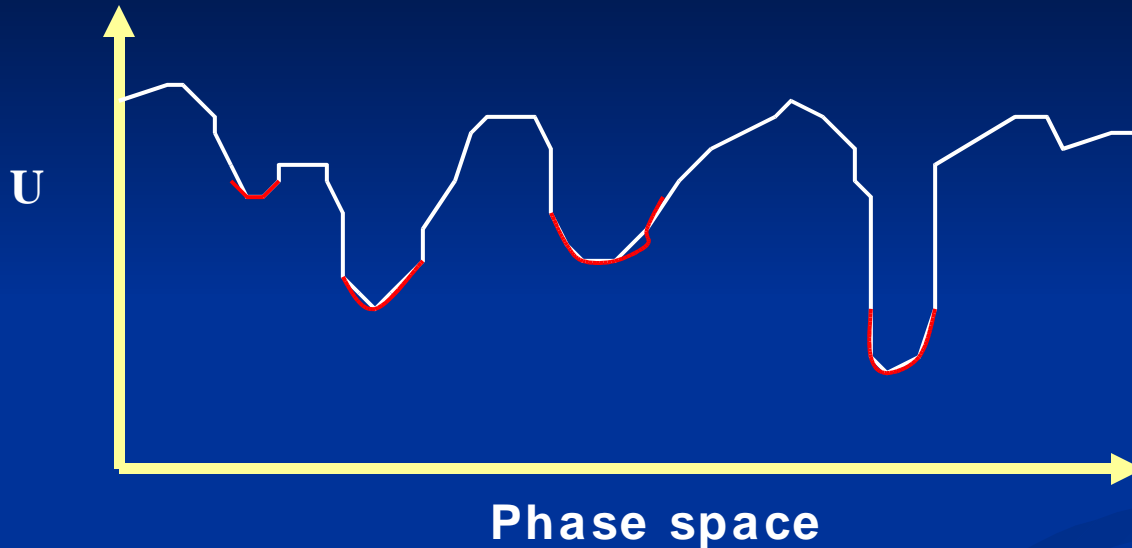


$$\Delta G_{A B} - \Delta G_{C B} - \Delta G_{A C} = 0$$

- Deviations are a measure of lack of convergence

None of the methods ensure your results are truly converged...

Why is it so difficult ?



- In docking, we only care about the single lowest energy point in phase space and the search algorithms can cut any corner to get there
- But in statistical physics, you need to know the **probability of each low energy point** in phase

Very difficult to explore quickly phase space and simultaneously obtain a good estimate of the probability of the low energy states

Current scope of free energy calculations:

Biotin/ Streptavidin

- COX-2. COX-1
- Trypsin
- HIV-1 protease
- P38

(...)

We need

- Better algorithms to run more quickly
- Better forcefields to get more accurate results
- Better sampling methods to get more precise results
- Better ways to simulate different ligands to deal with a more diverse set of compounds

Typical error in a “successful” free energy simulation = 1 kcal.mol⁻¹ (e.g., a 5-6 fold difference in IC₅₀ at room temperature)

Questions....

