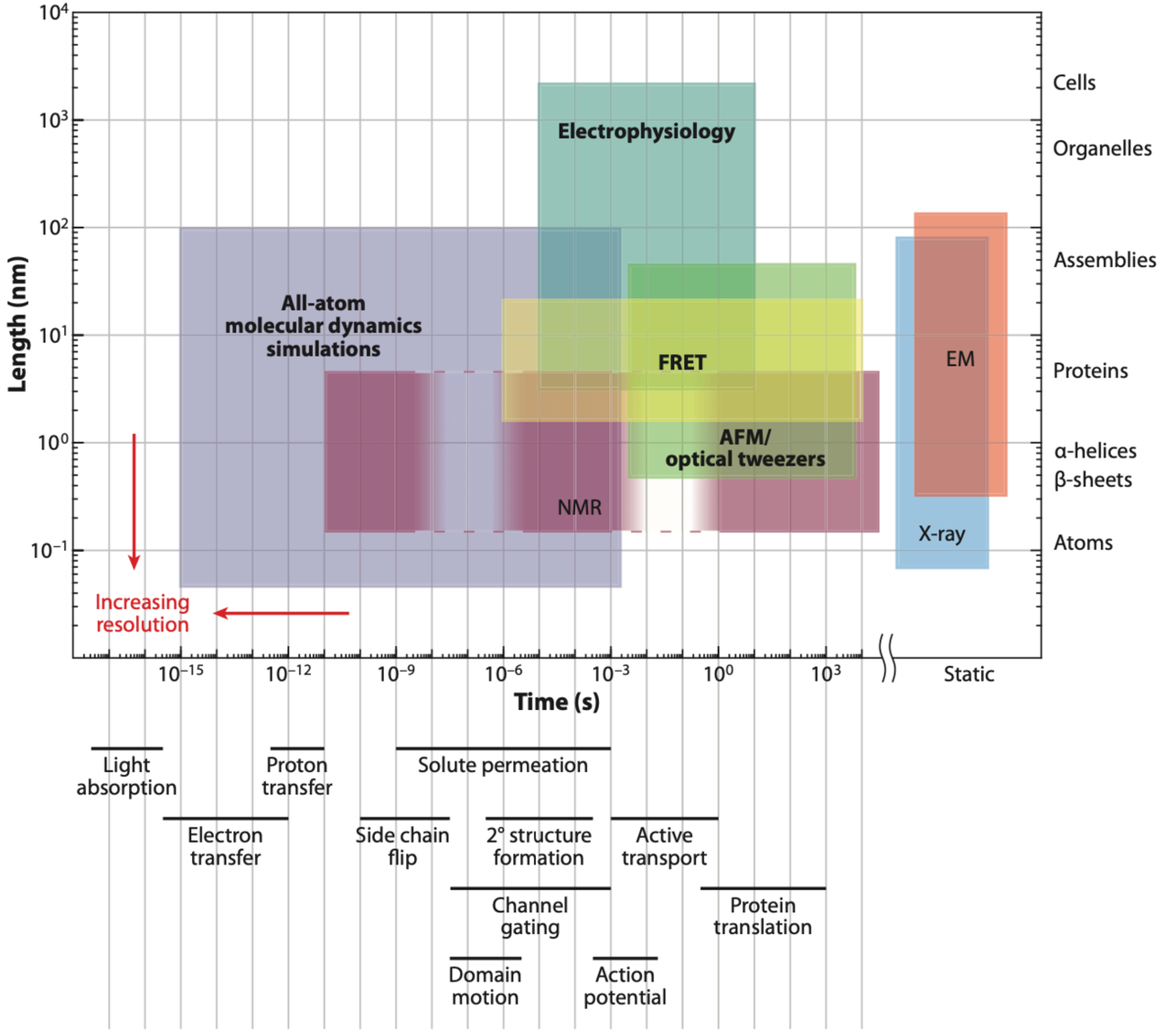
UNIVERSITÀ DEGLI STUDI DI MILANO

A mechano-statistical perspective of biomolecules in motion: towards building an in-silico microscope

Carlo Camilloni

Different techniques focus on different time scales:

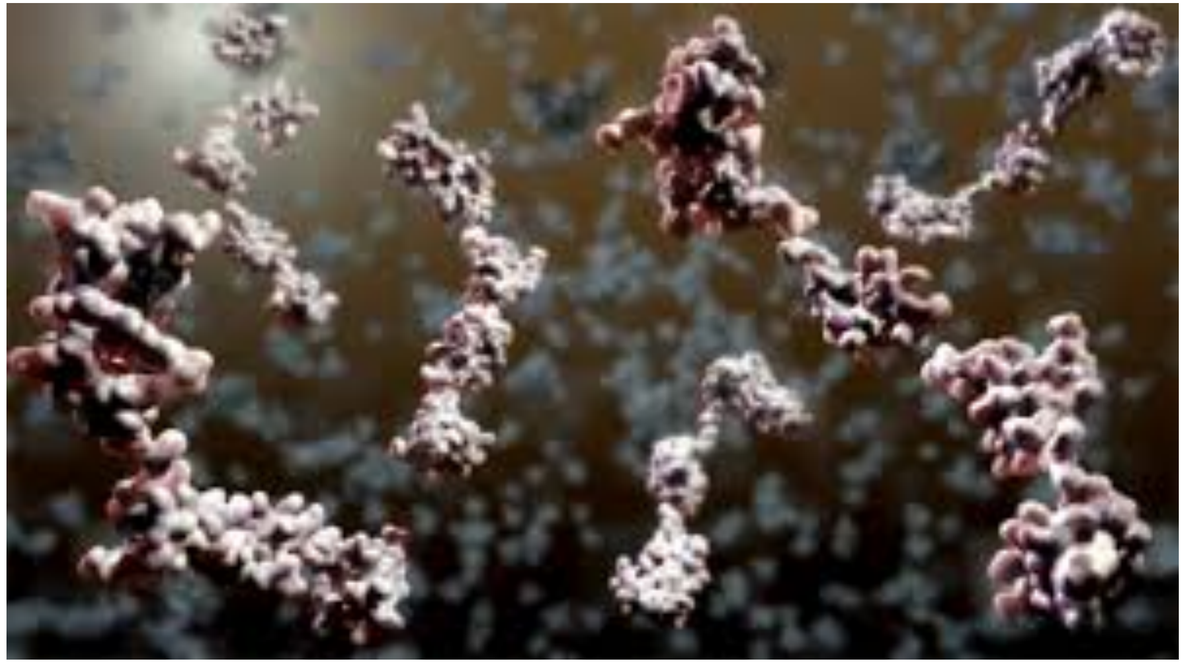
Different experimental techniques can explore processes at different spatial and time resolution, there is a region that can only be covered by computational methods, only NMR can get close to it, but its results are hard to interpret



Why a Statical and Mechanical perspective

Our observations will handle with many molecules, generally many copies of the same molecule, that can:

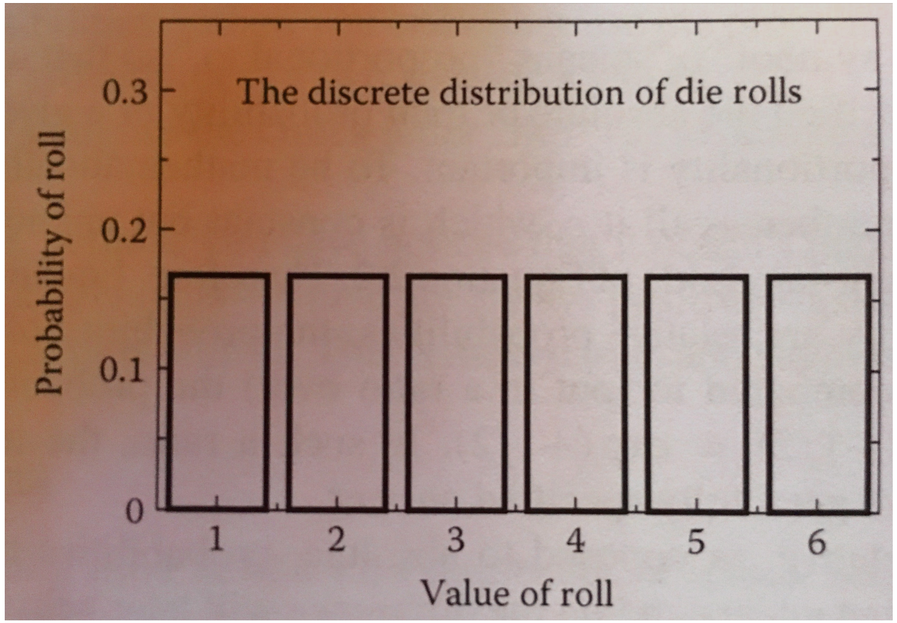
- change their shape;
- move in 3D;
- have interactions;



To describe all these we need to know how do they move and interact (mechanics) and how to handle large numbers (statistics)

A stochastic picture for molecules in motion: probabilities

Stochastic means randomly determined, events happen by chance, by observing many events one can learn **what is the probability of the different outcomes** that is one can learn their **probability distribution**. If some outcome is more likely than another one could think in terms of energy that that outcome is favourable in **energy**.



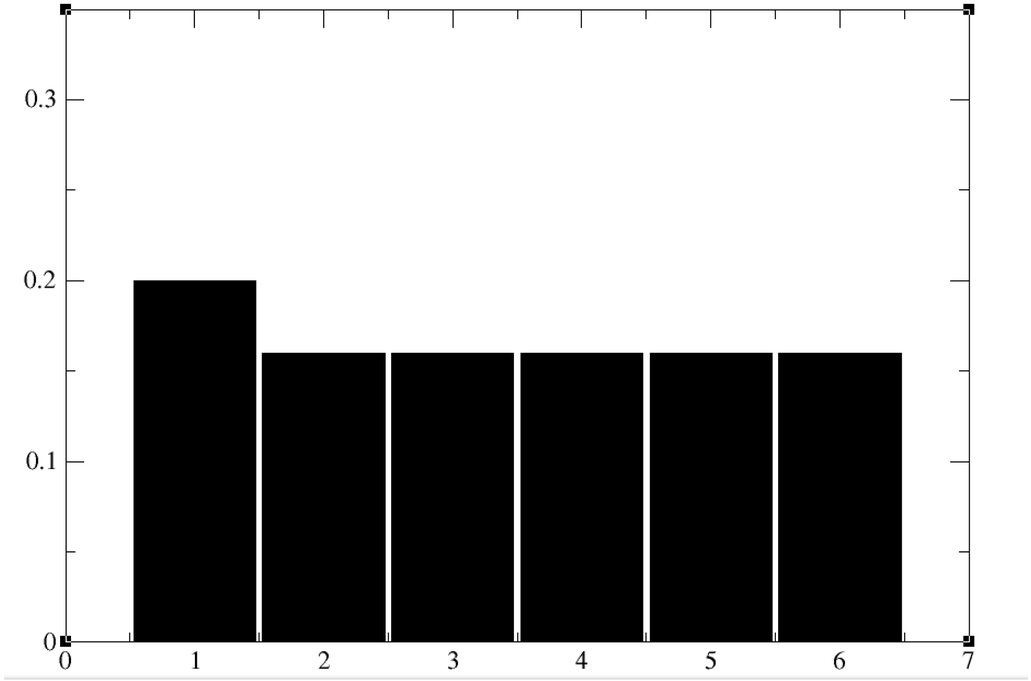
p(j)=1/6 with j=1 to 6 **Discrete**

$$\sum_{j=1}^{6} p(j) = 1$$

Normalised
When all possibilities are accounted for the probability is 100%

A stochastic picture for molecules in motion: biases

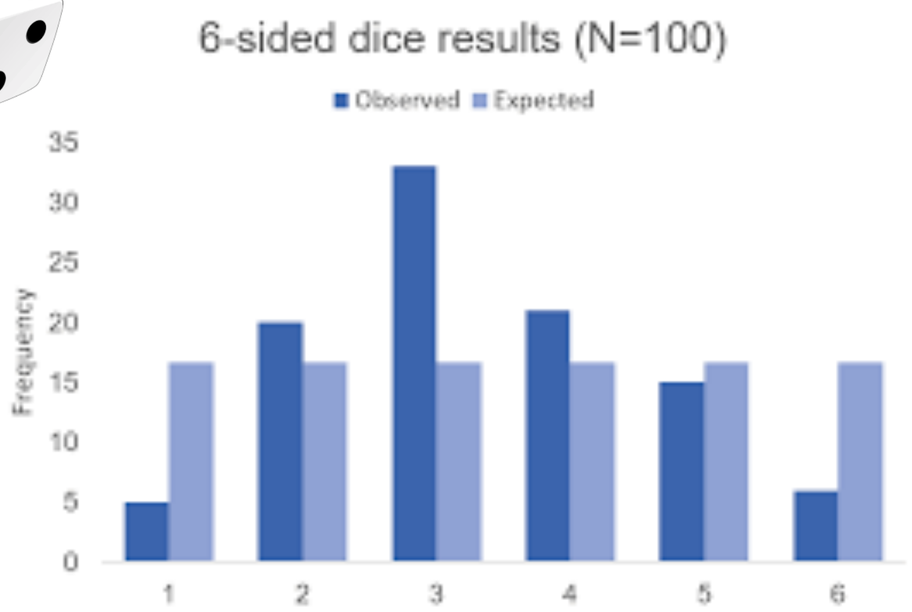
If your dice is biased towards a given number, let's say 1, then your histogram will be higher at 1. In terms of energy you can think that the state 1 is energetically more favourable.



$$\sum_{j=1}^{6} p(j) = 1$$

Normalised
When all possibilities are accounted for the probability is 100%

Probabilities are estimated using histograms: the problem of sampling relative and absolute probabilities



Stochastic events are estimated by histograms, where events are counted, these can deviate from the ideal expected behaviour due to random and systematic errors.

Absolute probability needs to have informations on all outcomes, relative probabilities are just frequency ratios:

Absolute probability: p(i) = count(i)/(sum_k(count(k)))
Relative probability: p(j)/p(i) = count(j)/count(i)

Ligand binding: a molecular example

Here energy and probability are intuitively linked, furthermore, events happen on some **time scale**, so one would like to find a relationship between probability, energy and time scales of events.

$$R + L \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} RL$$

In a ligand-receptor binding process we can picture the process, at least in some cases, as a probability for R and L to be unbound vs a probability to be bound (RL) and consider the time scale of the process in terms of rates of binding (on) and unbinding (off).

The k_{off} rate is the frequency of unbinding (s^{-1}) so that $k_{off}[RL]$ is the number of unbinding events per second.

The k_{on} represent instead two processes (1) collision and (2) complexation. So this is the collision frequency times the complexation probability, this means that binding depends on concentration, indeed k_{on} is measured in (s^{-1} M^{-1}). $k_{on}[R][L]$ is the number of binding events per second, while $k_{on}[L]$ is the frequency of binding for a given concentration of ligand (excess of ligand).

A stochastic picture: ligand binding

$$R + L \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} RL$$

If we start from a solution of free [R] and [L] we start forming RL with rate k_{on}, once we have some [RL] this can also dissociated with rate k_{off} , after some time we will be at <u>equilibrium</u> that is that the concentration of [R], [L] and [RL] will not change anymore

$$K_d \equiv \frac{k_{off}}{k_{on}} = \frac{[R][L]}{[RL]} = C^0 \frac{p_{unbound}}{p_{bound}}$$

The **dissociation constant** is defined as the ratio between k_{off} and k_{on}, it is measured in (M), where lower values means stronger binding (i.e. lower concentrations of ligand are enough to bind)

$$C^0 = 1\ M$$

$$\Delta G_{\text{binding}} = k_{\text{B}} T \ln\left(\frac{K_{\text{d}}}{C^0}\right) = k_{\text{B}} T \ln\left(\frac{k_{\text{off}}}{k_{\text{on}} C^0}\right)$$

$$= k_B T \ln\left(\frac{p_{unbound}}{p_{bound}}\right)$$

This is the binding free-energy that is the amount of work that can be extracted or that must be provided for the binding reaction to happen, and you see is related to the rates as well as to the probabilities

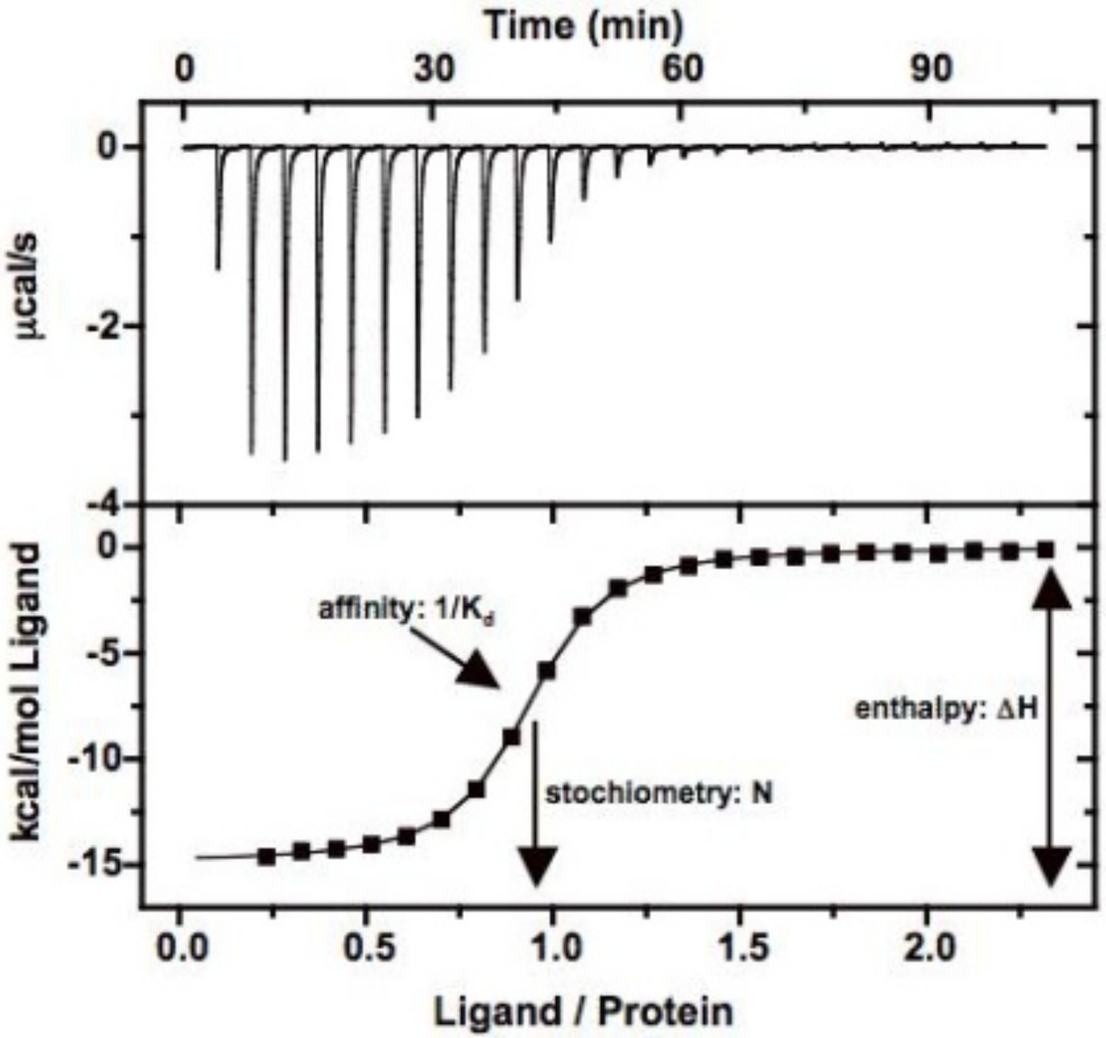
From experiments: Energy perspective: Isothermal Titration Calorimetry

Temperature is kept constant in some volume for:
- buffer alone
- buffer with protein and increased concentration of ligand

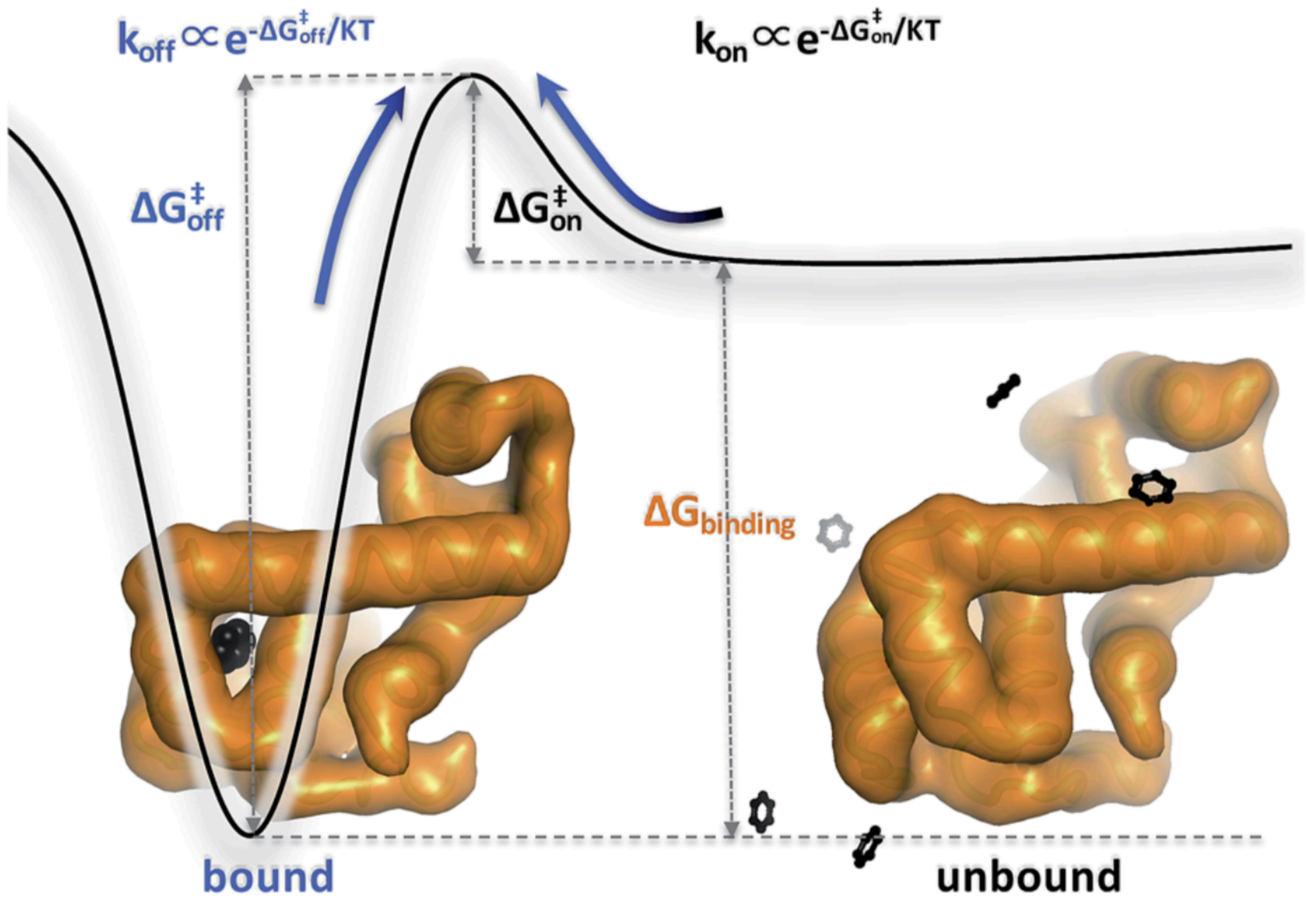
The amount of heat needed (or produced) to keep the temperature constant is measured

$$Q = [RL]\Delta HV$$

From the variation of Q and the concentration [LR] it is possible to recover ΔH and Kd



Binding is more complex than that not? Towards linking microscopic configurations to macroscopic processes



The macroscopic “bound” and “unbound” states can be realised by a very large number of microscopic configuration, of the receptor, of the ligand, but also of the solvent around them, if we could see them we should be then able to link them.

Free Energy

$$\Delta G = \Delta H - T\Delta S$$

We have seen that the free energy is the effective energy that gives the population of a state (e.g. bound/unbound). Macroscopically is the available work the system can perform, i.e. the maximum work you can extract from the system.

The enthalpy is the average energy of a state including the average fluctuations of the volume occupied at a given pressure. Macroscopically is the amount of heat given or absorbed by a reaction.

The entropy is the difference between the two and corresponds to the number of microscopic realisation of a macrostate (for example we can think of odds/even numbers in a dice and 1,3,5-2,4,6 being the corresponding microstates, in this case the entropy of the two macrostates is the same)

A stochastic picture: protein folding

$$U \underset{k_u}{\overset{k_f}{\rightleftharpoons}} F$$

The k_u rate is the frequency of unfolding (s^{-1}) so that $k_u[P]$ is the number of unfolding events per second.

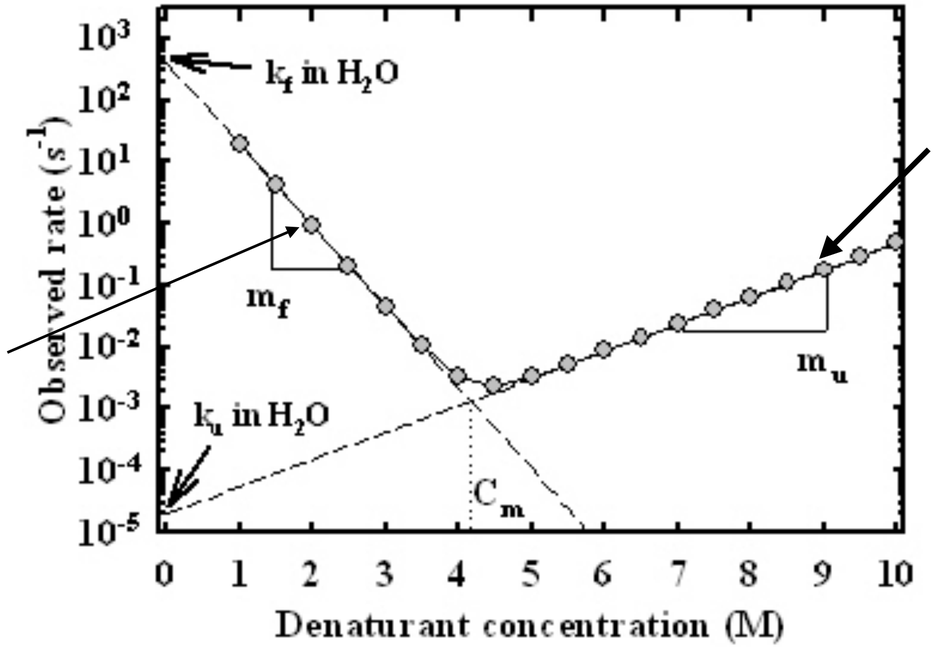
The k_f rate is the frequency of folding (s^{-1}) so that $k_f[P]$ is the number of folding events per second.

Their ratio define the $K_D=k_u/k_f$. This can used to calculate the difference in free energy:

$$\Delta G_{Eq} = RT \ln K_D$$

The two rates can be measured by changing the concentration of denaturants by a stopped-flow and monitoring fluorescence.

Protein in high-conc of denaturant is mixed with buffer to a target final low concentration of denaturant



Protein in buffer is mixed with denaturant to a target final high concentration of denaturant

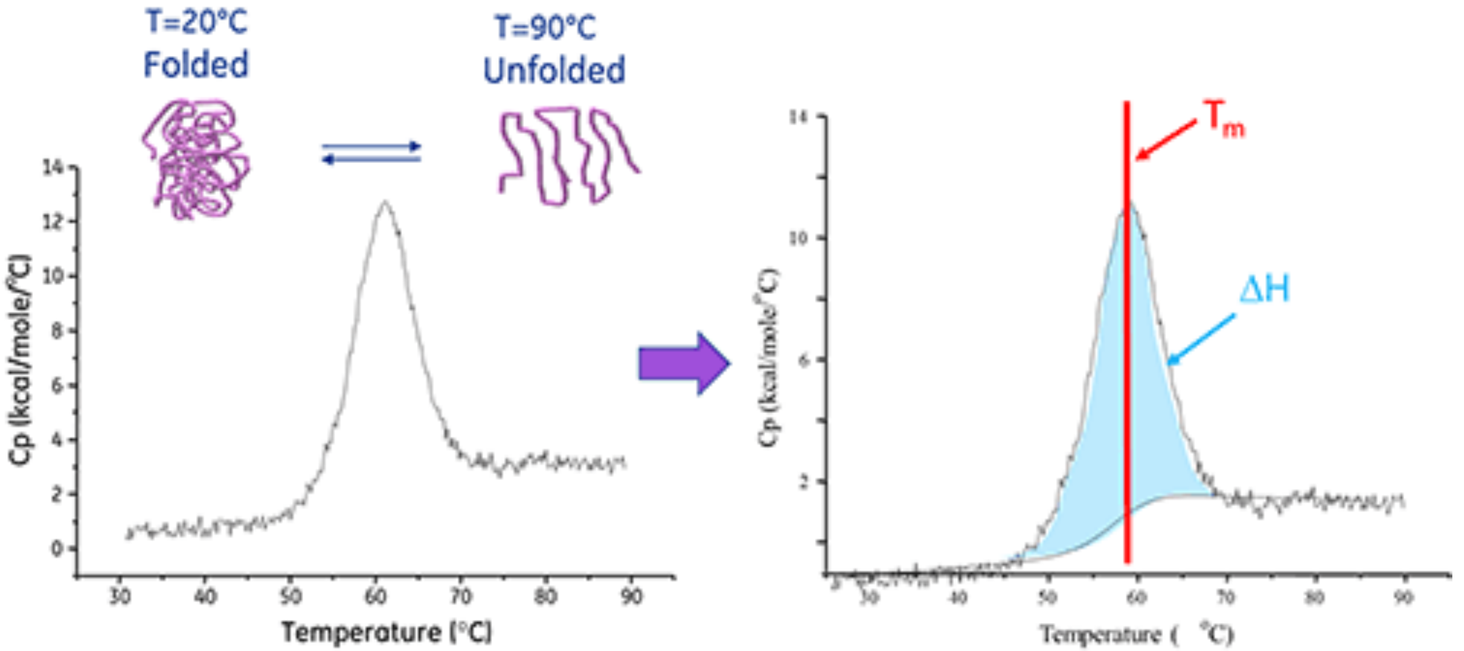
At high C all the protein unfold $k_{obs}=k_u$ at low C all the protein fold so $k_{obs}=k_f$

Energy perspective: Differential Scanning Calorimetry

Temperature is increased linearly in time:
- To buffer alone
- To buffer with protein

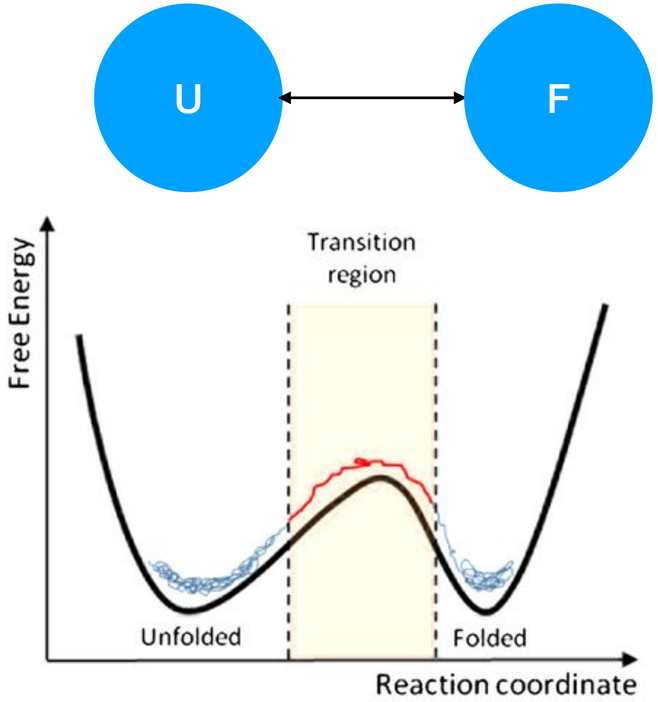
The amount of heat needed is different, by subtracting one gets the heat needed to increase the temperature of the protein.

Curve Cp vs temperature

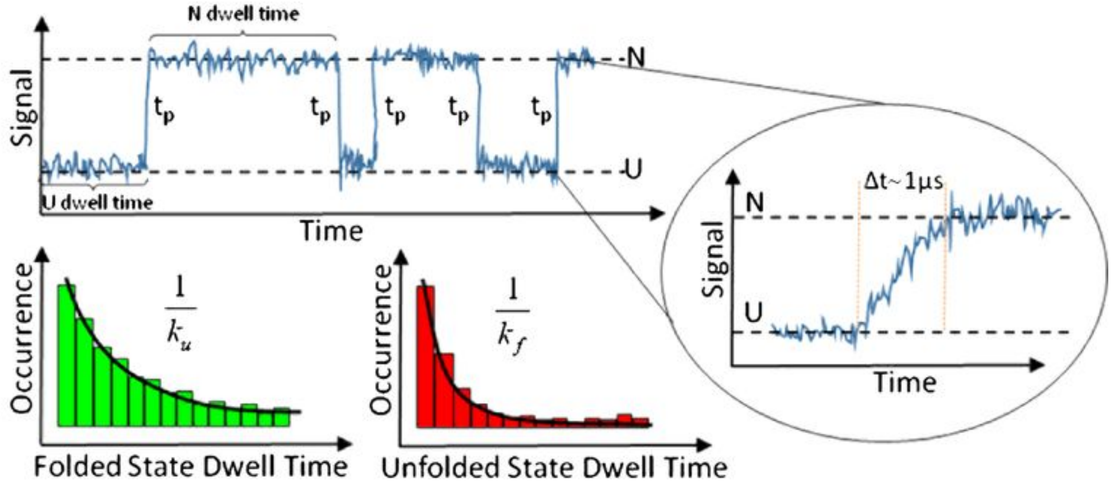


From the integral of Cp and it is possible to derive the Enthalpy difference.

Probability perspective: Single Molecule techniques (FRET, ..)



Where the balance between the two major macro states is determined by temperature



At room temperature the folded state can be ~99% populated, essentially suggesting

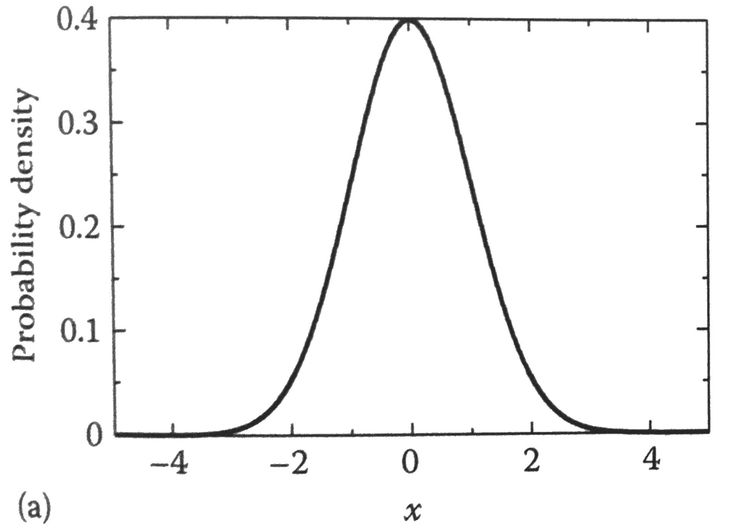
$$< O >= \sum_i p_i O_i \simeq O_f \approx ? O_{structure}$$

But one should not forget that a macro-state is still a collection of many micro-states

Continuos Probability Distributions

$$p_G(x) \propto e^{-\frac{(x-<x>)^2}{2\sigma^2}}$$

Gaussian



Density, they are in units^{-1}

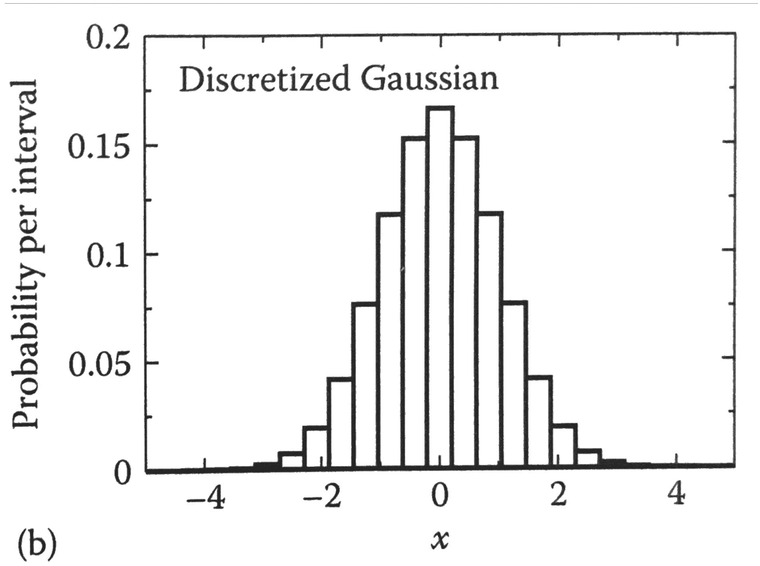
Even if we know only proportionality this can be enough to know make comparisons: the probability of p(0)/p(1). This is clear if you think about histograms.

$$\int_{-\infty}^{+\infty} e^{-\frac{(x-<x>)^2}{2\sigma^2}} dx = \sqrt{2\pi}\sigma$$

$$\rho_G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-<x>)^2}{2\sigma^2}}$$

Probability distributions are normalised

To calculate a probability one multiples for small interval:

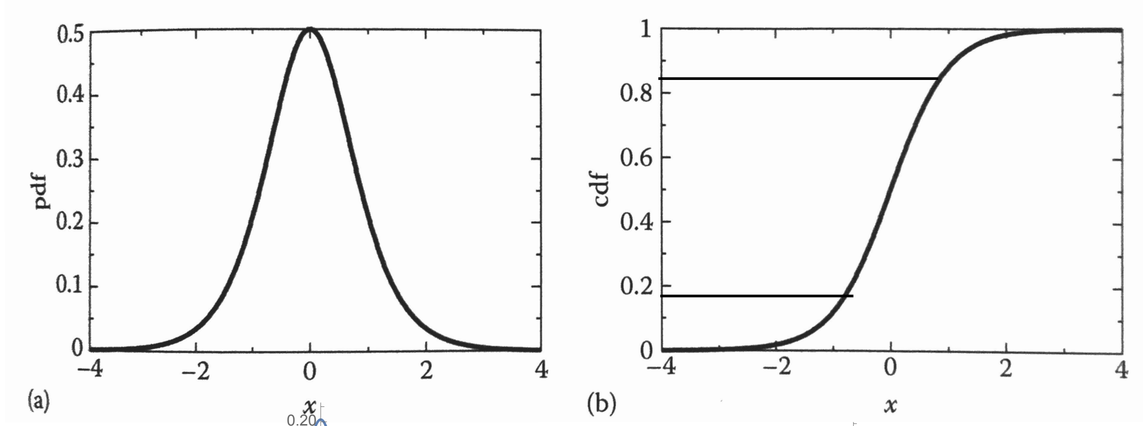


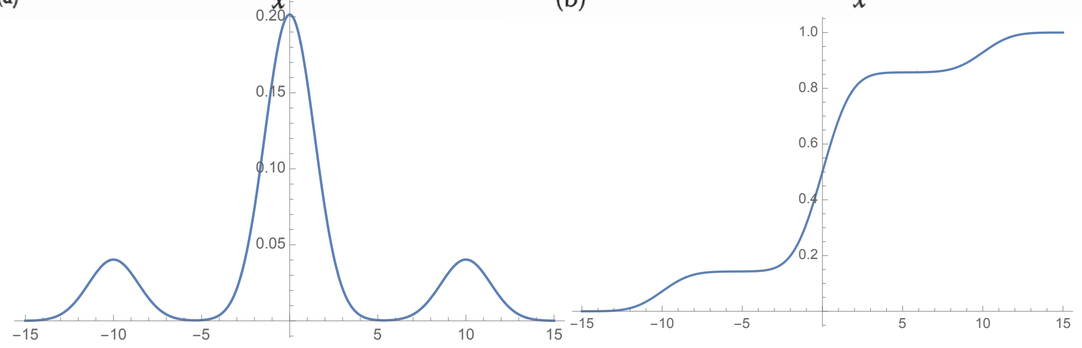
If you want a continuous probability distribution from a discrete sampling you still need to calculate the integral (i.e. the area)

Cumulative distribution function (CDF)

$$cdf(x) = \int_{-\infty}^{x} pdf(x')dx'$$

Is the amount of probability from the lower limit to x





Probability distributions can be multimodal

CDF are easier to describe the data:

- up to -1 there is only 10% of the data
- Between -1 and 1 there the 80% of the data

Sampling is the action of estimating a probability distributions. One takes elements and make histograms.

In an experiment data are distributed following the unknown distribution and you just count them. How do you generate numbers from a probability distribution? Using the CDF:

- take a random number y between 0 and 1
- take the x value corresponding to such y from the CDF

Average, Standard Deviation, Standard Error of the Mean

The definition of average is $< f(x) >=< f >= \int dx f(x)\rho(x)$

Where we have x representing an outcome of something, with probability density ρ(x), and we want to know the average of a function of the outcome *f(x). For example, if we just want to know the average outcome f(x)=x.*

If ρ(x) is not normalised, i.e. the integral of ρ(x) is not 1, or we want a weighted average

$$< f >= \frac{\int dx f(x)w(x)}{\int dx w(x)}$$

Here we can think at w(x) as the number of counts in the histogram for the outcome x and dx is the width of the histogram

Average, Standard Deviation, Standard Error of the Mean

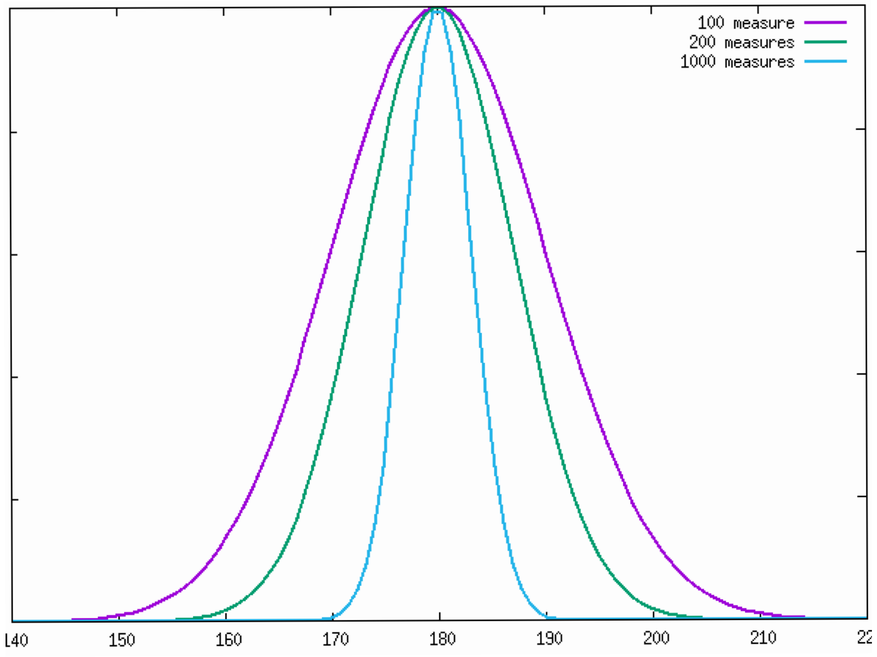
In practice we **estimate** averages using sums: $< f > \doteq \frac{1}{N}\sum_{i=1}^{N} f(x_i)$

Variance and Standard deviation: $\sigma^2 = var(f) =< (f - < f >)^2 >= \int dx (f(x) - < f >)\rho(x) \doteq \frac{1}{N-1}\sum_{i=1}^{N}(f(x_i) - < f >)^2$

Tell us about the width of the distribution ρ(x).

Standard error: $\text{Std-err} = \sqrt{\frac{var(f)}{N}} = \frac{\sigma}{\sqrt{N}} = \sqrt{\frac{1}{N^2 - N}\sum_{i=1}^{N}(f(x_i) - < f >)^2}$

Tell us about **accuracy of our estimate of the average,** the idea is that our average value that we estimate from multiple observations has a resolution with the shape of a Gaussian with the width of the standard error.

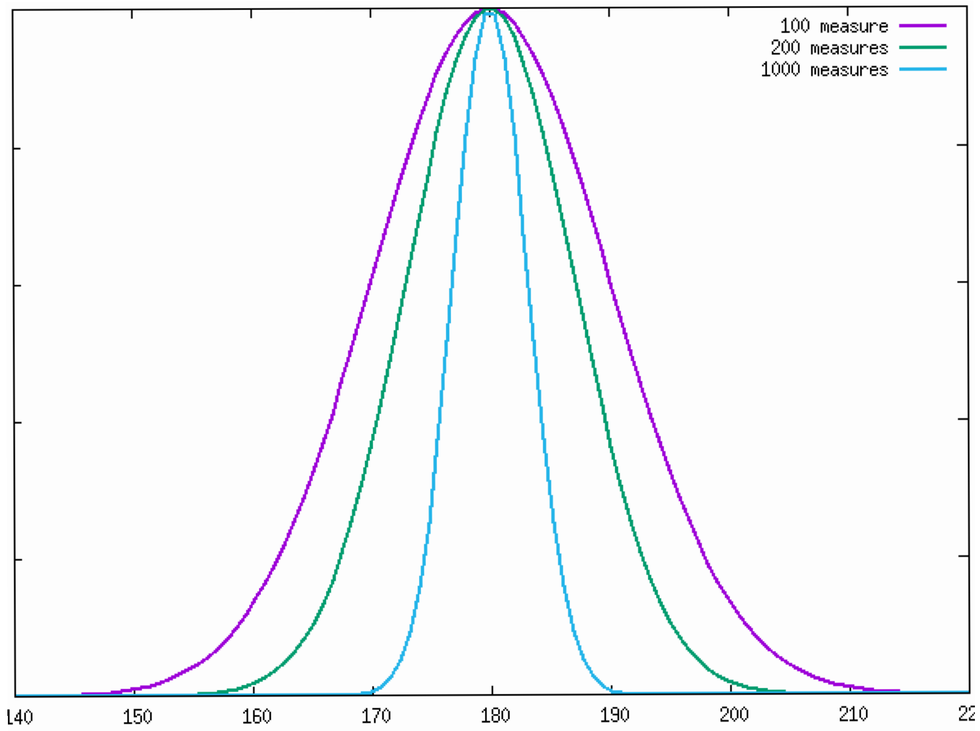


Average, Standard Deviation, Standard Error of the Mean

$$< f > \doteq \frac{1}{N}\sum_{i=1}^{N} f(x_i) \qquad \sigma^2 = var(f) = < (f - < f >)^2 > = \int dx (f(x) - < f >)\rho(x) \doteq \frac{1}{N-1}\sum_{i=1}^{N}(f(x_i) - < f >)^2$$

$$\text{Std-err} = \sqrt{\frac{var(f)}{N}} = \frac{\sigma}{\sqrt{N}} = \sqrt{\frac{1}{N^2 - N}\sum_{i=1}^{N}(f(x_i) - < f >)^2}$$

The average of a random process is always distributed as a gaussian of width std-err and so you can set confidence intervals: (±1std-err is 68%, ±2std-err 95%, ±3std-err 99%, …). This is called the theorem of the central limit.



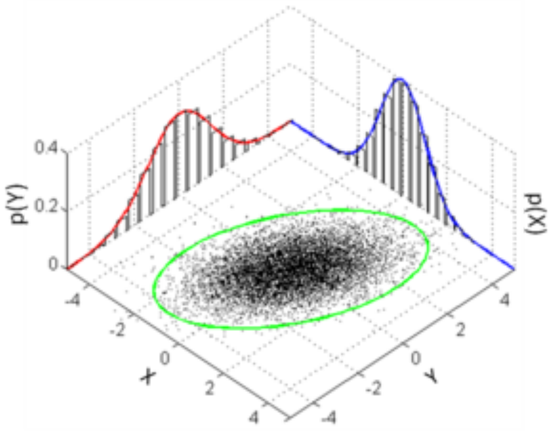
Multidimensional distribution function: projection and correlation

Probability distributions can be more than 1D. For example the probability distribution of a molecule around a receptor may be studied in terms of (x,y,z) coordinates of the center-of-mass, in this case would be 3D.

$\int \rho(x,y)dxdy = 1$ Normalisation

How can we show a multi dimensional distribution in less dimensions? By projecting it:

$\rho(x) = \int \rho(x,y)dy$ Projection/Marginalisation



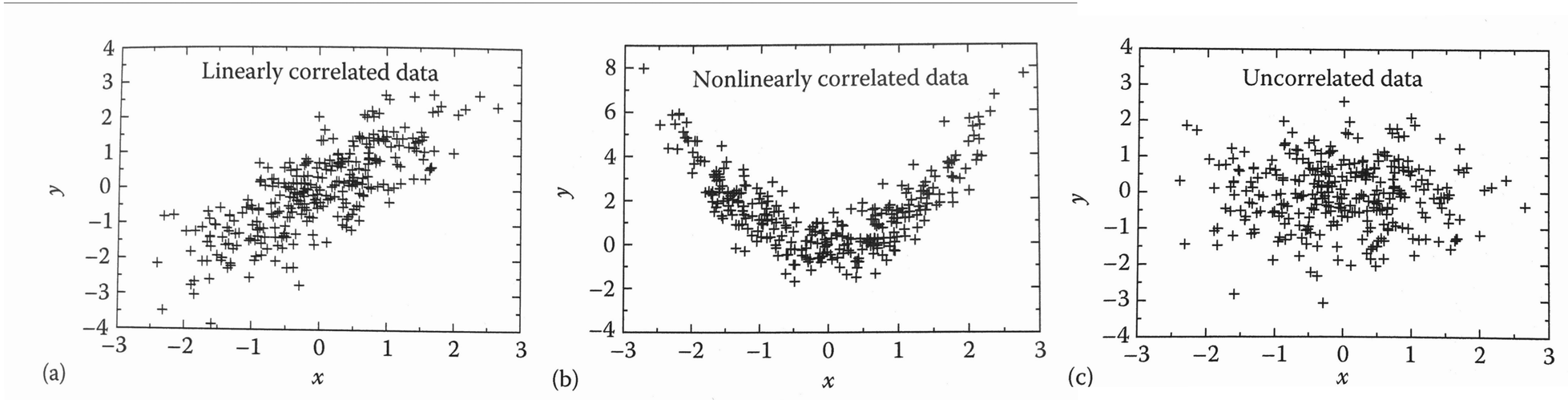
$$< x >= \int dxdyx\rho(x,y)$$

$$< y >= \int dxdyy\rho(x,y)$$

$$< xy > \int dxdyxy\rho(x,y)$$

By looking at the data can you think of a projection that is more important to understand the data?

Multidimensional distribution function: correlation



$$< x >= \int dxdyx\rho(x,y)$$

$$< y >= \int dxdyy\rho(x,y)$$

$$< xy > \int dxdyxy\rho(x,y)$$

If two variables are independent than

$$\rho(x,y) = \rho(x)\rho(y) \quad \textbf{So}$$

$$< xy >=< x >< y >$$

So we have correlation if

$$< xy > - < x >< y >\neq 0$$

Pearson correlation coefficient:

$$r = \frac{< xy > - < x >< y >}{\sigma_x \sigma_y}$$

Statistical Mechanics: the Boltzman distribution gives the probability to observe a molecule in a given conformation

What is the connection between statistics and mechanics?

That is the Boltzmann equation:

$$pdf(x,v) \equiv \rho(x,v) \propto \exp\left[\frac{-E(x,v)}{k_B T}\right] = \exp\left[\frac{-U(x)}{k_B T}\right] \exp\left[\frac{-K(v)}{k_B T}\right]$$

Let's look at this pdf in more detail:

$$pdf(x) \equiv \rho(x) \propto \exp\left[\frac{-U(x)}{k_B T}\right]$$

$$pdf(v) \equiv \rho(v) \propto \exp\left[\frac{-K(v)}{k_B T}\right] = \exp\left[\frac{-\frac{1}{2}mv^2}{k_B T}\right]$$

Conformations and velocities are independent. This means that we can study them separately. The distribution of the velocities does not affect the distribution of the configurations. Furthermore the distribution of the velocity is Gaussian, so it is easy to integrate it.

So for all practical purposes we can just ignore the velocity: $pdf(x) \equiv \rho(x) \propto \exp\left[\frac{-U(x)}{k_B T}\right]$

This gives a link between the energy of a conformation and the probability of observing it

Statistical Mechanics: Ergodicity

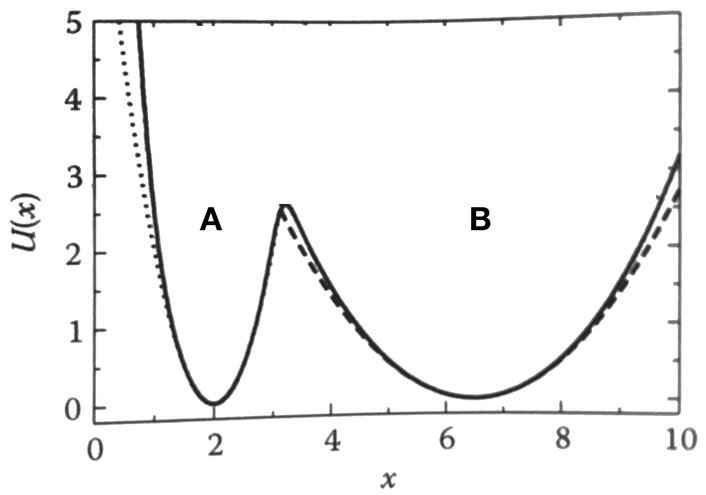
1. **A single molecule (at constant temperature) will move through all the possible conformations at random due to the collisions with the solvent. Each conformation is gonna be populated accordingly to the energy and temperature (the interaction with environment).**
2. **If at a given time one stops the movie of many many copies of the same system in the same condition the picture taken will display conformations according to the same statistics.**

The collection of conformations in both cases (that is the sampling) is usually called an ensemble of conformations. This ensemble is characterised first of all by the conditions in which it has been acquired (constant temperature, pressure or volume, constant number of particles, ecc)

States, Probabilities and Free-Energies

State: is a collection of configurations (microstate), ideally belonging to the same potential energy basin.

A simple 1D potential energy



Here we can visually define two states A and B for the two basins, e.g. all the conformations belonging to A and to B

$$p_A = \int_{V_A} \rho(x)dx \propto \int_{V_A} \exp[-U(x)/k_BT]dx$$

$$p_B = \int_{V_B} \rho(x)dx \propto \int_{V_B} \exp[-U(x)/k_BT]dx$$

The Free-Energy is the "effective" energy of a state, i.e. the energy that will give the same probability

$$\frac{p_A}{p_B} = \frac{\int_{V_A} \exp[-U(x)/k_BT]dx}{\int_{V_B} \exp[-U(x)/k_BT]dx} \equiv \frac{\exp(-F_A/k_BT)}{\exp(-F_B/k_BT)}$$

$$F_i \propto -k_BT \ln\left(\int_{V_i} \exp[-U(x)/k_BT]dx\right)$$

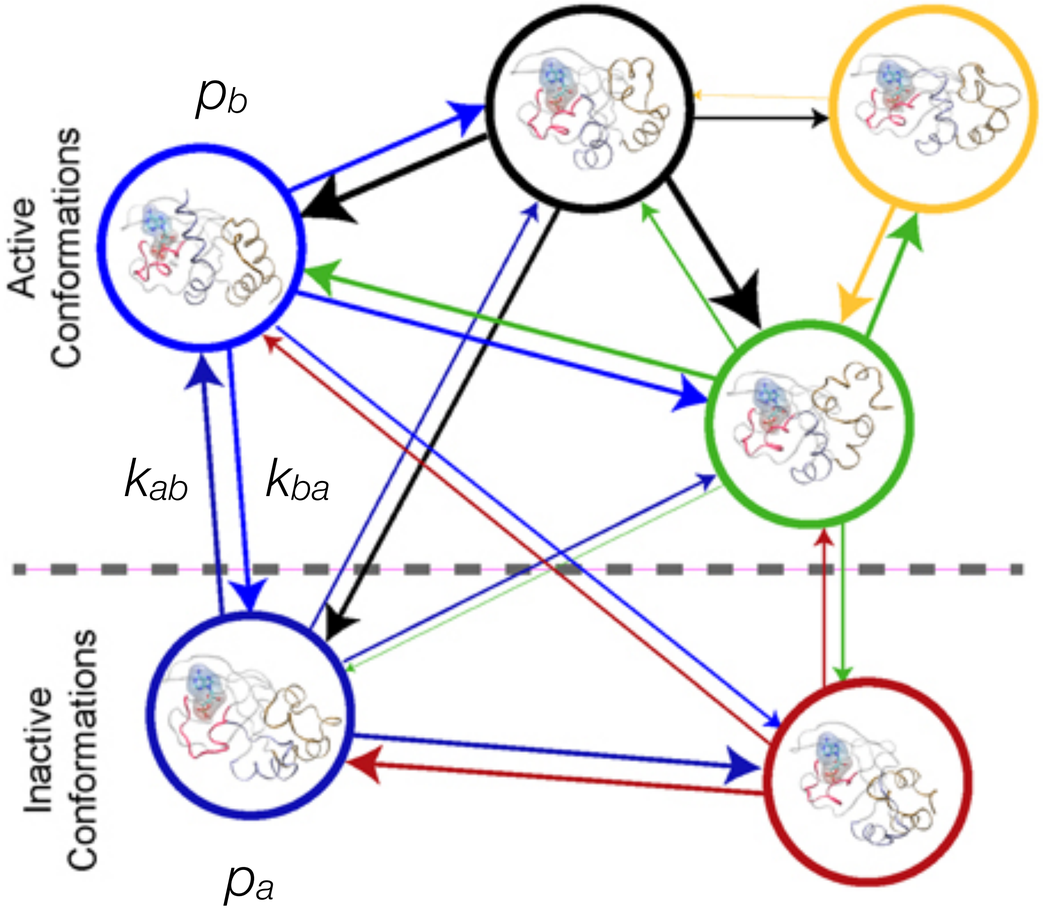
Summary: relative probabilities, energy and time

Microscopic processes are guided by **energy**, between two microscopic state their relative probability is associated exponentially to their energy difference.

Macroscopic processes (made of many microscopic one) are guided by **free-energy**, between two macroscopic state their relative probability is associated exponentially to their free-energy difference, the number of microscopic states making a macroscopic state is the entropy of the macroscopic state.

Macroscopic Processes happen in a given average time that is associated the free-energy barrier separating them

Summary: A stochastic picture of molecules in solution



Macrostate populations are associated with free energy differences:

$$p_i \propto \exp\left(-\frac{G_i}{RT}\right)$$

Exchange rates are associated with free energy barriers:

$$k_{ij} = k_0 \exp\left(-\frac{G_{ij}^{\sharp}}{RT}\right)$$

Energy for these processes is provided by thermal energy, so essentially by collision with water molecules

An equilibrium bulk experiment ideally capture the contribution from all the states:

$$< O >= \sum_i p_i O_i = \frac{\sum_i O_i \exp\left(\frac{-G_i}{RT}\right)}{\sum_i \exp\left(\frac{-G_i}{RT}\right)}$$

Microscopically:

$$G_i = -RT \ln \int \exp\left(-\frac{U(r)}{RT}\right) \delta(i - i(r)) dr$$

The common theme of Structural Bioinformatics: sampling and energy (or score)

In the next lectures we will see how the problem of SAMPLING and the problem of the ENERGY or SCORING FUNCTION, are interconnected and always present. These are the problem to cope with when doing MD simulation, structure prediction, docking, protein design, …

You can think at the next lectures as to specific strategies to cope with these two problems when trying to address specific problems. All the algorithms we will see are related to them (MD, MC, EM, ML/AI, …)

Ideally, what should a computational microscope do?

- Observe the time evolution of molecules at high spatial and time resolution
- Observe them for very long time scales
- Be aware of their energy
- Be able to set different experimental conditions (Temperature, Pressure, solution conditions)
- Be accurate
- Be interpretable, that is find suitable macrostates to compare with experiments
- …