Physique Statistique de Biomacromolecules

Paolo De Los Rios
Laboratoire de Biophysique Statistique
Institut de Théorie des Phénomènes Physiques
Ecole Polytechnique Fédérale de Lausanne

Semestre d’hiver 2004-2005

PROGRAM OF THE COURSE

1. **Theory of Random walks**

   The simplest model of polymers and a cornerstone of statistical physics

2. **Properties of Polymers**

   The basic biological macromolecules (DNA and proteins)

3. **Proteins**

   The workhorses of all organisms

A BRIEF HISTORY OF THE RANDOM WALK (Brownian motion)

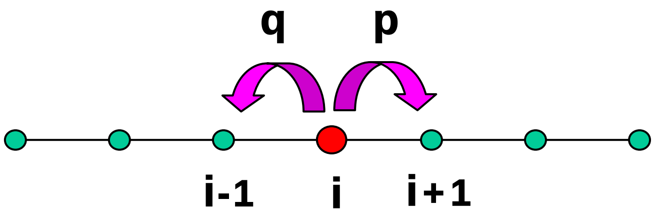
1827: ***Robert Brown*** **(a botanist) sees under his microscope that a pollen grain in water is moving all the time in a random way. By repeating the experience with different particles and liquids, he discovers that this movement is a universal property of the particle+liquid system.**

1905: ***Albert Einstein*** **uses microscopic principles to find the mathematical laws governing Brownian motion**

1905-10: ***Smoluchowski*** **and** ***Langevin*****, independently from Einstein, obtain equivalent descriptions of Brownian motion**

Random walk on a one-dimensional lattice

The Brownian particle can jump to the right with probability p and to the left with probability q



$p+q=1$

If the particle starts from i=0, what is the probability to find it at site n after N steps?

Probability to find the particle at n after N steps

$$P(n,N)=\frac{\#\,of\ \ favorable\ cases}{\#\,of\ \ possible\ cases}$$

The total number of possible cases is simply 2^N

To arrive in n after N steps, the random particle must go N_R times to the right and N_L times to the left with the double constraint

$$N_R + N_L = N$$

$$N_R - N_L = n$$

These constraints give

$$N_R = \frac{N + n}{2} \qquad N_L = \frac{N - n}{2}$$

So the number of favorable cases is given by the number of distinguishable ways to mix N_R and N_L steps

$$\#\, \boldsymbol{of \;\; favorable\; cases} = \frac{\boldsymbol{N}\,!}{\boldsymbol{N_R}\,!\left(\boldsymbol{N}-\boldsymbol{N_R}\right)!}$$

And finally we obtain

$$\boldsymbol{P}(\boldsymbol{n},\boldsymbol{N}) = \frac{1}{2^{N}}\frac{\boldsymbol{N}\,!}{\left(\dfrac{\boldsymbol{N}+\boldsymbol{n}}{2}\right)!\left(\dfrac{\boldsymbol{N}-\boldsymbol{n}}{2}\right)!}$$

We use Stirling’s approximation to obtain an asymptotic expression for P(n,N) in the limit N»n. The result is

$$\boldsymbol{P}(\boldsymbol{n},\boldsymbol{N}) \approx \frac{2}{\sqrt{2\pi \boldsymbol{N}}}\,\boldsymbol{e}^{-\frac{\boldsymbol{n}^2}{2\boldsymbol{N}}}$$

Is a gaussian distribution surprising?

The position after **N** steps is the sum of **N** random variables equal to **$x_i = \pm 1$** with equal probability.

$$n(N) = \sum_{i=1}^{N} x_i$$

For large values of N the Central Limit Theorem tells us that the distribution of n(N) tends to a gaussian distribution of square variance N

Continuous space and time description of Brownian motion

Previously we have focused on counting possible paths, one by one; now we look at all of them at once.

We need a completely probabilistic description of the random walk!

We want a relation between P(n,N+1) and P(n,N)

Master equation for P(n,N)

The probability of the particle being in n at time N+1 is related to its probability of being in a different site n’ at time N, such that it could jump from n’ to n in one time step.

If the probability of jumping at a distance k (positive or negative) is p(k), then we have

$$\boldsymbol{P}(\boldsymbol{n}, \boldsymbol{N}+1) = \sum_{\boldsymbol{k}=-\infty}^{+\infty} \boldsymbol{p}(\boldsymbol{k})\, \boldsymbol{P}(\boldsymbol{n}-\boldsymbol{k}, \boldsymbol{N})$$

We look at the simple case

$$p(k) = \frac{1}{2}\delta_{k,+1} + \frac{1}{2}\delta_{k,-1}$$

and we obtain the master equation

$$P(n, N+1) = \frac{1}{2}P(n-1, N) + \frac{1}{2}P(n+1, N)$$

Going to the continuum limit

$$N \rightarrow t \qquad N+1 \rightarrow t+\Delta t$$

$$n \rightarrow x \qquad n+1 \rightarrow x+\Delta x$$

and therefore

$$P(x,t+\Delta t) = \frac{1}{2} P(x-\Delta x,t) + \frac{1}{2} P(x+\Delta x,t)$$

After some simple algebra, the continuum time and space limit is

$$\lim_{\Delta x \to 0,\, \Delta t \to 0} \frac{P(x, t+\Delta t) - P(x,t)}{\Delta t} =$$

$$= \lim_{\Delta x \to 0,\, \Delta t \to 0} \left[\frac{\Delta x^2}{2\,\Delta t} \left(\frac{P(x-\Delta x, t) - 2P(x,t) + P(x+\Delta x, t)}{\Delta x^2} \right) \right]$$

The limit is then

$$\frac{\partial}{\partial t} P(x,t) = D \frac{\partial^2}{\partial x^2} P(x,t)$$

This is the
DIFFUSION EQUATION

And the diffusion constant *D* is

$$D = \lim_{\Delta x \to 0,\, \Delta t \to 0} \frac{\Delta x^2}{2\Delta t}$$

This is the first appearance of a special space-time relation

$$\Delta \boldsymbol{x} \;\square\; \Delta \boldsymbol{t}^{1/2}$$

Without this relation the diffusion constant *D* would be either null or infinite

The solution of the diffusion equation is obtained by Fourier transform

$$\frac{\partial}{\partial t}\overset{\square}{P}(k,t) = -Dk^{2}\,\overset{\square}{P}(k,t)$$

The solution is $\overset{\square}{P}(k,t) = \overset{\square}{P}(k,0)e^{-Dk^{2}t}$

Where $\overset{\square}{P}(k,0)$ is the transform of the initial condition

Taking the anti-transform yields

$$P(x,t)=\frac{1}{\sqrt{4\pi Dt}}\int_{-\infty}^{+\infty}P(x',0)e^{-(x-x')^2/4Dt}dx'$$

And if the initial condition is $P(x,0)=\delta(x)$

$$P(x,t)=\frac{1}{\sqrt{4\pi Dt}}e^{-\frac{x^2}{4Dt}}$$

Second time we find the special space-time relation

$$P(x,t)=\frac{1}{\sqrt{4\pi Dt}}e^{-\frac{x^2}{4Dt}}$$

Space scales as the square root of time!!!

The variance of the distribution is

$$< x^2 > = 2Dt$$

The variance grows linearly in time. In particular the diffusion constant is defined, from experiments, using the above formula.

The diffusion equation can be derived also in *d* dimensions

$$P(\vec{x},t+\Delta t)=\frac{1}{2d}\sum_{i=1}^{d}\left[P(\vec{x}+\Delta x_i\vec{e}_i,t)+P(\vec{x}-\Delta x_i\vec{e}_i,t)\right]$$

Where $\vec{e}_i$ is the unit vector in the i_{th} direction.

The known trick for the continuum limit yields

$$\frac{\partial}{\partial t}P(\vec{x},t)=\sum_{i=1}^{d}D_i\frac{\partial^2}{\partial x_i^2}P(\vec{x},t) \qquad D_i=\lim_{\substack{\Delta x_i\to 0\\ \Delta t\to 0}}\frac{\Delta x_i^2}{2d\Delta t}$$

The diffusion equation is not necessarily isotropic!

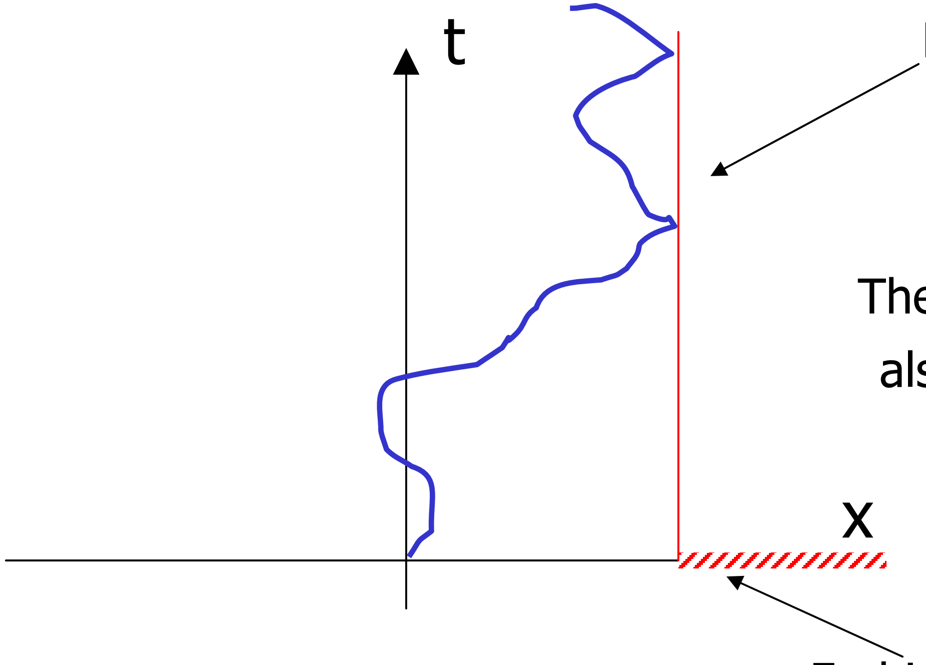
The solution of the high dimensional diffusion equation is then

$$P(\vec{x},t)=\frac{1}{(4\pi t)^{d/2}\left(\prod_{i=1}^{d} D_i\right)^{1/2}}\,e^{-\sum_{i=1}^{d}\frac{x_i^2}{4D_i t}}$$

The relation between space and time does not change!!!

Random walk in bounded spaces

The simplest example of a bounded space is the half line

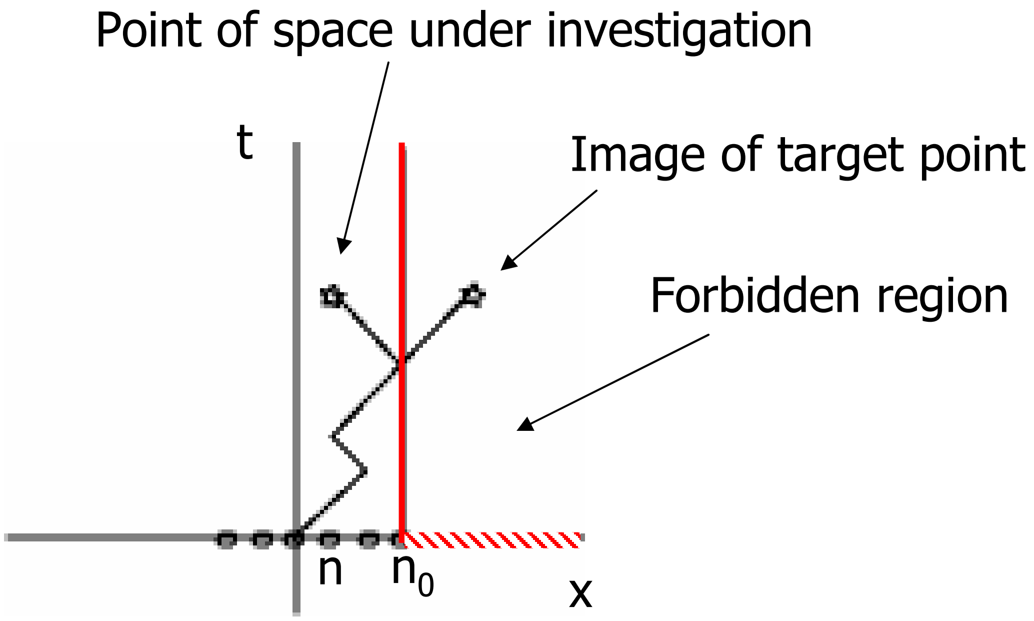


Reflective boundary

The boundary could also be absorbing

Forbidden region

Random walk on a bounded lattice



Point of space under investigation

Image of target point

Forbidden region

What is the probability P(n,N) of being at n after N time steps, knowing that there is reflecting/absorbing wall in n_0?

Reflecting wall

If the wall is reflecting, all paths that would go the image point end up in the target site.

The image point is located at 2n_0-n, and therefore

$$\boldsymbol{P}(\boldsymbol{n}, N; \boldsymbol{n}_0) = \boldsymbol{P}(\boldsymbol{n}, N) + \boldsymbol{P}(2\boldsymbol{n}_0 - \boldsymbol{n}, N)$$

Absorbing wall

If the wall is absorbing, all paths that would go the image point are destroyed.

The image point is located at 2n_0-n, and therefore

$$\boldsymbol{P}(\boldsymbol{n},\boldsymbol{N};\boldsymbol{n}_0)=\boldsymbol{P}(\boldsymbol{n},\boldsymbol{N})-\boldsymbol{P}(2\boldsymbol{n}_0-\boldsymbol{n},\boldsymbol{N})$$

We can easily go to the continuum

Reflecting wall

$$P(x,t;x_0) = P(x,t) + P(2x_0 - x,t)$$

Absorbing wall

$$P(x,t;x_0) = P(x,t) - P(2x_0 - x,t)$$

The principle of the image is very useful in one dimension and with a simple wall. In more complex geometries, and in higher dimensions it cannot be applied.

What do we learn then from the image trick that can be applied in other contexts?

Since our aim is to find a solution for the diffusion equation, we look for the boundary conditions imposed by the principle of the image

Reflecting wall

$$P_r(x,t;x_0) = P(x,t) + P(2x_0 - x,t)$$

$$\Downarrow$$

$$\partial_x P_r(x,t;x_0)\,|_{x_0} = \partial_x P(x,t)\,|_{x_0} + \partial_x P(2x_0 - x,t)\,|_{x_0} = 0$$

But $\partial_x P_r(x,t;x_0)\,|_{x_0}$ is the current of probability across the wall. If it is null it means that there there is no probability that a particle crosses the wall, in perfect agreement with physical intuition for a reflecting barrier.

We can then generalize this boundary condition to arbitrary geometry and dimension

$$\hat{\boldsymbol{n}} \cdot \vec{\nabla} \boldsymbol{P}(\vec{\boldsymbol{x}}, \boldsymbol{t})\,|_{\vec{x} \in \boldsymbol{boundary}} = 0$$

Where $\hat{\boldsymbol{n}}$ is the unitary vector normal to the boundary

Absorbing wall

$$P_r(x,t;x_0) = P(x,t) - P(2x_0 - x,t)$$

$$\Downarrow$$

$$P_r(x_0,t;x_0) = P(x_0,t) - P(2x_0 - x_0,t) = 0$$

The probability to find the particle on the wall is therefore zero, which is physically intuitive since as the particle touches the wall, it disappears, so that the probability to find it there is null

We can then generalize this boundary condition to arbitrary geometry and dimension

$$P(\vec{x},t)\,\big|_{\vec{x}\in boundary}=0$$

Normalization

The particle can be found only on the left of the wall.

Reflecting wall

$$\int_{-\infty}^{x_0} P(x,t;x_0)dx = \int_{-\infty}^{x_0} P(x,t)dx + \int_{-\infty}^{x_0} P(2x_0 - x,t)dx =$$

$$= \int_{-\infty}^{x_0} P(x,t)dx + \int_{x_0}^{+\infty} P(x,t)dx = 1$$

The probability is conserved, compatibly with
the absence of a probability current across the wall.

Absorbing wall

$$\int_{-\infty}^{x_0} P(x,t;x_0)dx = \int_{-\infty}^{x_0} P(x,t)dx - \int_{-\infty}^{x_0} P(2x_0 - x,t)dx =$$

$$= \int_{-\infty}^{x_0} P(x,t)dx - \int_{x_0}^{+\infty} P(x,t)dx < 1$$

Since the probability to find the particle left of the wall is smaller than 1, this implies that the particle has been absorbed before time t with the probability

$$P_{death}(t) = 1 - \int_{-\infty}^{x_0} P(x,t;x_0)dx$$

What is the total probability of death?

For large time, does the particle always die?

$$P_{survival}(t)=\frac{1}{\sqrt{4\pi Dt}}\int_{-\infty}^{x_0}dx\left[e^{-\frac{x^2}{4Dt}}-e^{-\frac{(x-2x_0)^2}{4Dt}}\right]=$$

$$=\frac{2}{\sqrt{4\pi Dt}}\int_{-\infty}^{x_0}e^{-\frac{x^2}{4Dt}}dx-1=\frac{2}{\sqrt{2\pi}}\int_{-\infty}^{x_0/\sqrt{2Dt}}e^{-\frac{y^2}{2}}dy-1$$

This expression goes to 0 in the large time limit!

The probability of having died before time t is the complement of the survival probability

$$P_{death}(t) = 1 - P_{survival}(t) = 2 - \frac{2}{\sqrt{2\pi}} \int_{-\infty}^{x_0/\sqrt{2Dt}} e^{-\frac{y^2}{2}} dy$$

Since the particle can die just once, if we define $P_{die}(t)dt$ the probability to die between t and t+dt, then we can also write

$$P_{death}(t) = \int_0^t P_{die}(t')dt'$$

The probability to die can be obtained as

$$P_{die}(t) = \frac{d}{dt} P_{death}(t) = \frac{x_0}{\sqrt{4\pi D}\ \ t^{3/2}} e^{-\frac{x_0^2}{4Dt}}$$

$P_{die}(t)$ is also the probability that the particle's life is *t*.

As a consequence, the average life of a particle is

$$\langle t \rangle = \int_0^\infty t \cdot P_{die}(t) dt = \frac{x_0}{\sqrt{4\pi D}} \int_0^\infty t^{-1/2} e^{-\frac{x_0^2}{4Dt}} dt = \infty$$

We have therefore the strange result that, although the particle dies with probability 1, its average life-time is infinite.

This might seem counterintuitive, but this is a typical problem with Levy distributions. Indeed $P_{die}(t)$ is a Levy distribution with infinite first and second moments.

Geometric picture

How can we visualize the “sure death” of the Brownian particle?

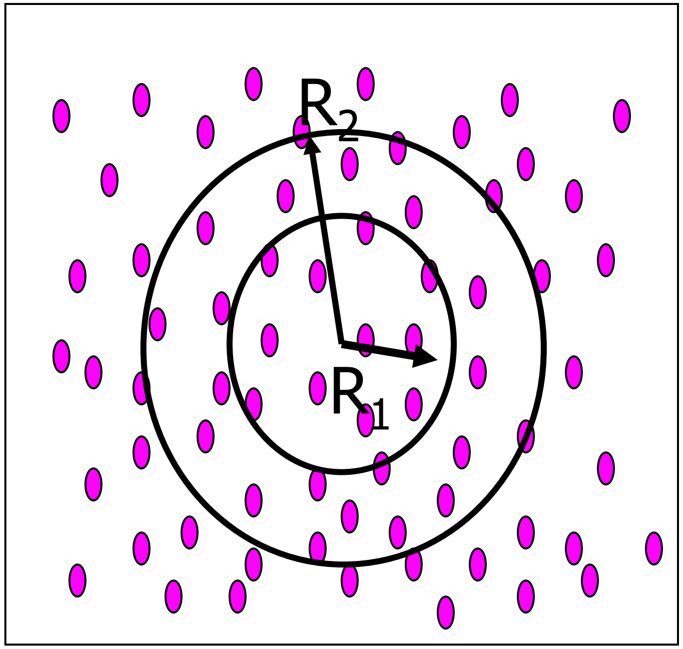
In order to understand why the random walk will meet a sure death we need a new concept: the Fractal Dimension of its trajectory.

The fractal dimension extends to non-integer values the usual concept of Euclidean dimension that we are used to.

We define sets of objects that are geometrically intermediates between lines and surfaces, and between surfaces and solids.

Hausdorff fractal dimension

How does the number of points within a circle of radius R grow with R?



If the distribution is uniform we expect that $N(R) \sim R^2$

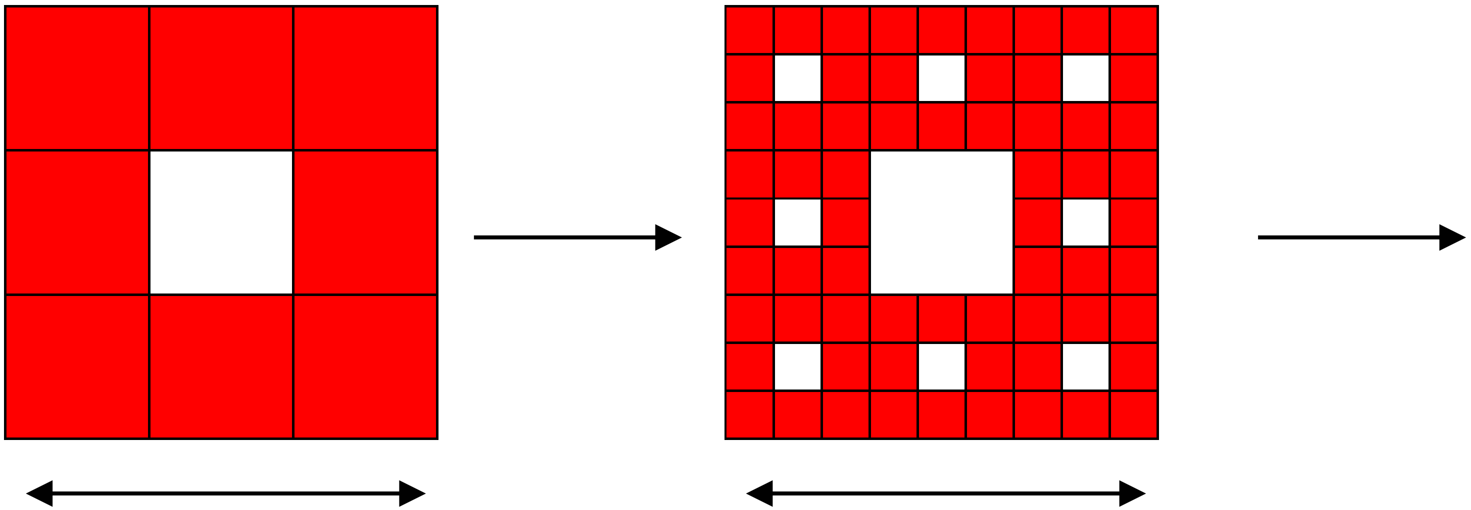
The points of the surface itself are distributed homogeneously: therefore the surface has an area $A(R) = \pi R^2$

It is important not to mix the dimension of the space where the collection of objects is embedded, and the dimension of the collection of objects!!!

In the previous example, the embedding space dimension is d=2. The Hausdorff dimension of the collection of objects, instead, is given by the scaling relation between their number and the linear dimension R. In the previous example we had

$$\boldsymbol{D_H} = 2$$

A simple (regular) fractal



and so on

L$_1$=3
N$_1$=8

L$_2$=9
N$_2$=64

$$N = L^{D_H}$$

$$D_H = \log_3 8 < 2$$

The fractal dimension of Brownian motion

Given a trajectory, how many of its points fall within distance R of the origin?

First we have to give an estimate of the linear dimension of the region that contains the random walk.

We define two vectors:

$$\vec{\boldsymbol{R}}_0 = \frac{1}{\boldsymbol{N}} \sum_{\boldsymbol{n}=1}^{\boldsymbol{N}} \vec{\boldsymbol{R}}_{\boldsymbol{n}} \qquad \longleftarrow \qquad \text{Center of mass}$$

$$\boldsymbol{R}_{\boldsymbol{g}}^2 = \frac{1}{\boldsymbol{N}} \sum_{\boldsymbol{m}=1}^{\boldsymbol{N}} (\vec{\boldsymbol{R}}_{\boldsymbol{m}} - \vec{\boldsymbol{R}}_0)^2 \qquad \longleftarrow \qquad \text{Radius of gyration}$$

We are of course interested in average quantities
(a single random walk is not very interesting)

$$\left\langle \vec{\boldsymbol{R}}_0 \right\rangle = \frac{1}{\boldsymbol{N}} \sum_{\boldsymbol{n}=1}^{\boldsymbol{N}} \left\langle \vec{\boldsymbol{R}}_{\boldsymbol{n}} \right\rangle = 0 \qquad \longleftarrow \quad \text{by symmetry}$$

$$\boldsymbol{R}_{\boldsymbol{g}}^2 = \frac{1}{\boldsymbol{N}} \sum_{\boldsymbol{m}=1}^{\boldsymbol{N}} \left\langle (\vec{\boldsymbol{R}}_{\boldsymbol{m}} - \vec{\boldsymbol{R}}_0)^2 \right\rangle = \frac{1}{\boldsymbol{N}} \sum_{\boldsymbol{m}=1}^{\boldsymbol{N}} \left(\left\langle \vec{\boldsymbol{R}}_{\boldsymbol{m}}^2 \right\rangle + \left\langle \vec{\boldsymbol{R}}_0^2 \right\rangle - 2 \left\langle \vec{\boldsymbol{R}}_{\boldsymbol{m}} \cdot \vec{\boldsymbol{R}}_0 \right\rangle \right)$$

The second and third terms in the last expression decrease as $1/N^2$ and $1/N$ with respect to the first, so that in the large N limit they can be neglected.

The radius of gyration can therefore be evaluated as

$$R_g = \sqrt{\frac{1}{N}\sum_{m=1}^{N}\left\langle \vec{R}_m^2 \right\rangle} \;\square\; \sqrt{\frac{\sigma^2}{N}\sum_{m=1}^{N} m} \;\square\; \sigma N^{1/2}$$

(once again we find the key space-time relation typical of Brownian motion and of diffusion!!!)

So we find at last that N objects (the points of Brownian motion) are contained in a region of linear size $L=N^{1/2}$. The fractal dimension is

$$N \sim L^{D_H} = N^{D_H/2}$$

and therefore $D_H=2$

First important remark: the fractal dimension of Brownian motion is 2 independently of the dimension of the embedding space!!! (the space time relation does not depend on the embedding dimension)

Second remark: the fractal dimension of Brownian motion is 2 even in one dimension.

The path of a random walk is a two dimensional object with spherical symmetry! It’s a disordered object.

Simplest interpretation in two dimensions:
Since the path of a random walk is a two-dimensional structure, it covers the plane completely.
So, *a fortiori,* it also covers the one dimensional line completely. This means that it will pass for each and every point of the line, including the wall, and therefore death is certain.

This geometrical interpretation tells us that the random walk finds a particular point of the plane also with probability one, since it covers the plane completely!

In three dimensions, instead, the path of the random walk does not fill the environment, and it has therefore some hope to avoid death!!!

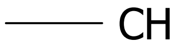
Polymers

Polymers are chains of units (monomers). Usually the monomers belong to the realm of organic chemistry (that is chemistry based on carbon), and understanding polymers of biological relevance will be our ultimate goal.

Polymers can be classified according to the type of units they are made of, and most of their chemical properties depends on the details of the sequence along the chain.

Yet, there are some general features that do not depend too much on the details of the chemistry.

Types of monomers



Polyethylene oxide

Poly(N-isopropylacrylamide)

Proteins (20 different types of aminoacids)

Polymers composed of a single type of monomers are called homopolymers, else they are called heteropolymers.

Heteropolymers come in different forms:
if the sequence of monomer species along the chain is random, they are called random heteropolymers; if the heterogeneity is regular (periodicity or others), they are called block copolymers.

Proteins are not random, but you couldn’t tell from the seqeunce!!!

The statistical properties of polymers are different from these of random walks.

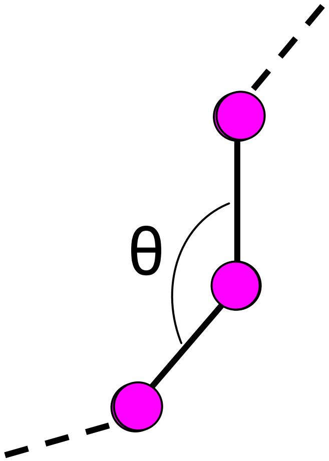
In particular the end-to-end distance (and the gyration radius) have a different scaling behavior

$\boldsymbol{R} \sim \boldsymbol{N}^{1/2}$ Random Walks

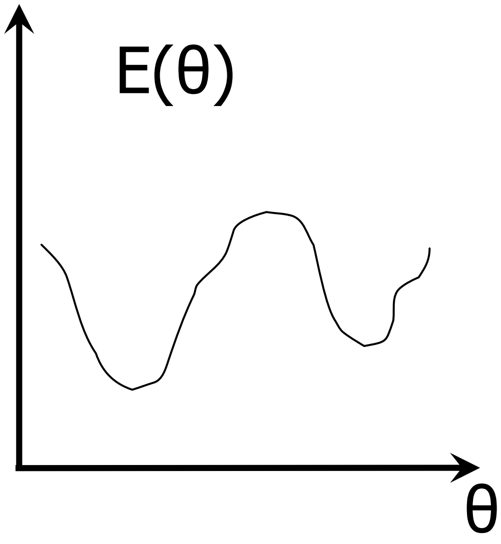
$\boldsymbol{R} \sim \boldsymbol{N}^{\nu} \quad \boldsymbol{\nu} > 1/2$ Polymers

Where does this difference come from?

Let’s start from unstructured polymers



Due to the bonding properties
of the carbon (or other) atoms,
not every angle is equally probable:
some angles cost less energy than others



If all angles were possible with the same probability (that is, they cost the same energy), then the polymer would correspond to a *Freely Jointed Chain* (FJC). The FJC corresponds to a random walk with steps all of the same length (this last approximation is due to the fact that usually the bonds along the backbone are “rigid”: it is difficult to stretch them).

Introducing a chain stiffness makes it more difficult to bend the polymer, and we can expect a larger end-to-end distance that for a random walk. Is it enough to explain the exponent difference?

The Worm-Like Chain (WLC)

We make the simple assumption that, due to the presence of some preferential angle between two successive segments, the average

$$\left\langle \vec{\boldsymbol{t}}_{\boldsymbol{i}} \cdot \vec{\boldsymbol{t}}_{\boldsymbol{i}+1} \right\rangle = \boldsymbol{a} \neq 0$$

Then we set out to evaluate the scaling of the gyration radius with the number of segments, N

We compute the radius of gyration of the WLC

$$\vec{\boldsymbol{R}}_M = \frac{1}{\boldsymbol{N}}\sum_{\boldsymbol{i}=1}^{\boldsymbol{N}}\left\langle \vec{\boldsymbol{R}}_{\boldsymbol{i}} \right\rangle = 0$$ by symmetry

$$\boldsymbol{R}_g^2 = \frac{1}{\boldsymbol{N}}\sum_{\boldsymbol{i}=1}^{\boldsymbol{N}}\left\langle \left(\vec{\boldsymbol{R}}_{\boldsymbol{i}} - \vec{\boldsymbol{R}}_M\right)^2 \right\rangle = \frac{1}{\boldsymbol{N}}\sum_{\boldsymbol{i}=1}^{\boldsymbol{N}}\left\langle \left(\vec{\boldsymbol{R}}_{\boldsymbol{i}} \cdot \vec{\boldsymbol{R}}_{\boldsymbol{i}}\right) \right\rangle =$$

$$= \frac{1}{\mathrm{N}}\sum_{\boldsymbol{i}=1}^{\boldsymbol{N}}\sum_{\boldsymbol{l}=1}^{\boldsymbol{i}}\sum_{\boldsymbol{k}=1}^{\boldsymbol{i}}\left\langle \vec{\boldsymbol{t}}_{\boldsymbol{k}} \cdot \vec{\boldsymbol{t}}_{\boldsymbol{l}} \right\rangle$$

We must compute $\left\langle \vec{t}_k \cdot \vec{t}_l \right\rangle$.

We know that $\left\langle \vec{t}_k \cdot \vec{t}_{k+1} \right\rangle = a \leq 1$

Then we can write

$$\left\langle \vec{t}^{\,T}_{k} \cdot \vec{t}_{k+2} \right\rangle = \left\langle \vec{t}^{\,T}_{k} \cdot \underline{\underline{I}} \cdot \vec{t}_{k+2} \right\rangle =$$

$$= \left\langle \vec{t}^{\,T}_{k} \cdot \left(\vec{t}_{k+1} \vec{t}^{\,T}_{k+1} + \vec{y}_{k+1} \vec{y}^{\,T}_{k+1} + \vec{z}_{k+1} \vec{z}^{\,T}_{k+1} \right) \cdot \vec{t}_{k+2} \right\rangle =$$

$$= \left\langle \vec{t}^{\,T}_{k} \cdot \vec{t}_{k+1} \vec{t}^{\,T}_{k+1} \cdot \vec{t}_{k+2} \right\rangle + \left\langle \vec{t}^{\,T}_{k} \cdot \vec{y}_{k+1} \vec{y}^{\,T}_{k+1} \cdot \vec{t}_{k+2} \right\rangle + \left\langle \vec{t}^{\,T}_{k} \cdot \vec{z}_{k+1} \vec{z}^{\,T}_{k+1} \cdot \vec{t}_{k+2} \right\rangle$$

In the previous expressions $\vec{y}_{k+1}$ and $\vec{z}_{k+1}$ are unit vectors orthogonal to $\vec{t}_{k+1}$ and the products are not correlated since the energy depends only on the products of consecutive vectors. Moreover, we can choose $\vec{z}_{k+1}$ to be orthogonal to both $\vec{t}_k$ and $\vec{t}_{k+1}$

$$\left\langle \vec{t}^{\,T}_{k} \cdot \vec{t}_{k+1} \vec{t}^{\,T}_{k+1} \cdot \vec{t}_{k+2} \right\rangle = \left\langle \vec{t}^{\,T}_{k} \cdot \vec{t}_{k+1} \right\rangle \left\langle \vec{t}^{\,T}_{k+1} \cdot \vec{t}_{k+2} \right\rangle = a^2$$

$$\left\langle \vec{t}^{\,T}_{k} \cdot \vec{y}_{k+1} \vec{y}^{\,T}_{k+1} \cdot \vec{t}_{k+2} \right\rangle = \left\langle \vec{t}^{\,T}_{k} \cdot \vec{y}_{k+1} \right\rangle \left\langle \vec{y}^{\,T}_{k+1} \cdot \vec{t}_{k+2} \right\rangle = b^2$$

Then we find

$$\left\langle \vec{t}_k \cdot \vec{t}_j \right\rangle = \sum_{l=0}^{j-k} \binom{j-k}{l} a^l\, b^{j-k-l} = (a+b)^{j-k}$$

$$R_g^2 = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} \sum_{k=1}^{i} \sum_{j=1}^{i} \left\langle \vec{t}_k \cdot \vec{t}_j \right\rangle = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{i} \sum_{j=1}^{i} (a+b)^{j-k} =$$

$$= \frac{2}{N} \sum_{i=1}^{N} \sum_{k=1}^{i-1} \sum_{j=k}^{i} (a+b)^{j-k} - \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{i} 1 \underset{N \gg 1}{=} \frac{1}{2} \frac{1+(a+b)}{1-(a+b)} N$$

So we have clearly that $\boldsymbol{R}_{\boldsymbol{g}}^{2} \propto \boldsymbol{N}$ and the WLC is not different from the simple FJC (that is, the random walk).

The ingredient that we have introduced to deal with the WLC is the local bending energy. It introduces correlations along the chain, but they decrease exponentially with the distance between the segments (*k>i*):

$$\left\langle \vec{\boldsymbol{t}}_{\boldsymbol{i}} \cdot \vec{\boldsymbol{t}}_{\boldsymbol{k}} \right\rangle = (\boldsymbol{a}+\boldsymbol{b})^{\boldsymbol{k}-\boldsymbol{i}} = \boldsymbol{e}^{(\boldsymbol{k}-\boldsymbol{i})\ln(\boldsymbol{a}+\boldsymbol{b})}$$

Where $\lambda = 1/\ln(\boldsymbol{a}+\boldsymbol{b})$ is the persistence length of the chain.

In conclusion, a random walk with an exponentially decreasing correlation along the chain is not enough to break the typical random walk result and the essence of the Central Limit Theorem.

This is due to the fact that, by a simple coarse graining of the chain where a number of steps of the order of λ is substituted by a single step, then consecutive steps would be uncorrelated (since they extend over a distance of the order of the correlation length) and the random walk result would follow.

To go beyond the random walk, then we need to introduce stronger correlations along the chain.

What kind of correlations do we need?

A possibility could be to introduce correlations that decay slower than an exponential: an algebraic decay would work because it is unaffected by the coarse graining. Yet there is no physical motivation able to justify such correlations.

Another possibility is to introduce correlations that are intrinsically non-local along the chain by asking that no pair of monomers shares the same region of space, no matter how far they are along the chain. It is a long range correlation along the chain that has a good physical motivation. It can, in principle, violate the Central Limit Theorem. Does it do it in reality?

How can we evaluate whether the mutual exclusion of monomers will be enough to find a result different from Brownian motion?

In d=1 it is clear that it works: once the polymer has chosen a direction, then it can only go straight on. Hence in d=1 $R_g \sim N$.

We can resort to the fractal dimension of the random walk to begin understanding what happens in d>1.

First we need a simple mathematical tool: the fractal dimension D_I of the intersection of two fractal objects of dimensions D_1 and D_2 in dimension d:

$$D_I = D_1 + D_2 - d$$

We can then ask what is the fractal dimension of a random walk with itself:

$$\boldsymbol{D_I} = 2 + 2 - \boldsymbol{d} = 4 - \boldsymbol{d}$$

How do we use it?

d=1 : D_I=3, that is, the number of intersections grows with the length of the polymer and and it will affect R_g, as we have seen;

$2 \leq d < 4$: $D_I > 0$, and intersections should still be important, changing the behavior from random walk;

$d \geq 0$: $D_I \leq 0$ (which means actually =0) and intersections should not play a role. As a consequence $R_g \sim N^{1/2}$.

The Method of Flory

We can try to understand a little better the behavior of polymers that satisfy the excluded volume constraint: they are called self-avoiding walks (SAW).

The method of Flory is a way to estimate the exponent ν in the expression $R_g \sim N^{\nu}$.

The important question is: what is the probability that the end-to-end distance of a SAW is a given R?

If the SAW was a simple random walk we would know the answer

$$P_0(R,N) = \frac{1}{\sqrt{2\pi\sigma^2 N}} e^{-\frac{R^2}{2\sigma^2 N}}$$

But we know that only a fraction of the random walks can actually be considered: those without self-intersections.

So, the true probability for a SAW is

$$\boldsymbol{P}(\boldsymbol{R},\boldsymbol{N}) = \boldsymbol{P}_0(\boldsymbol{R},\boldsymbol{N}) \cdot \boldsymbol{S}(\boldsymbol{R},\boldsymbol{N})$$

Where S(R,N) is the survival probability of random walks, that is the probability that they have no self-intersections.
The second step is then to evaluate S(R,N).
Actually, it is easier to evaluate the death probability 1-S(R,N), because it is simply the probability that there are intersections.
Flory found a “mean field” expression for that probability.

Estimate of 1-S(R,N)

The probability that a monomer is found in a given region of space dV (volume element in d dimensions) is simply

$$p_1 dV = \rho dV = \frac{N}{V} dV$$

where ρ=N/V is the monomer density inside a volume V.

The probability that TWO monomers fall into the same region of space is

$$p_2 dV = \underbrace{dV\,\frac{N}{V}}_{1st\ monomer}\ \underbrace{\int_v dv\,\frac{N}{V}}_{\substack{probability\ that\ the \\ 2nd\ monomer\ is\ in \\ the\ volume\ of\ the\ 1st}}$$

where *v* is the volume of a monomer. This expression, integrated over the whole volume, gives the death probability

$$1 - S(R,N) = v\frac{N^2}{V}$$

The volume V is the volume occupied by the polymer: a sphere of radius R_g in d dimensions: $V \sim R_g^d$

There is a further approximation: the density of monomers $\rho \sim 1/V$ means that monomers are uniformly distributed inside the volume. This is of course not true.

At last we have

$$S(R,N) = 1 - v\frac{N^2}{R_g^d} \underset{V\to\infty}{\approx} e^{-v\frac{N^2}{R_g^d}}$$

Putting everything together we finally have

$$\boldsymbol{P}(\boldsymbol{R}_g, \boldsymbol{N}) = \frac{1}{\sqrt{2\pi\sigma^2 \boldsymbol{N}}} \boldsymbol{e}^{-\left(\frac{\boldsymbol{R}_g^2}{2\sigma^2 \boldsymbol{N}} + \boldsymbol{v}\frac{\boldsymbol{N}^2}{\boldsymbol{R}_g^{\boldsymbol{d}}}\right)}$$

Attention!!! In all this we have freely exchanged end-to-end distance and gyration radius. This is legitimate, in the spirit of the Flory method, because they show the same scaling with respect to N, and because anyway the method is approximate.

The next step is to find the R_g for which $P(R_g,N)$ is maximal.

$P(R_g,N)$ is maximal when the argument of the exponential is maximal too.

$$-\frac{R_g}{N}+\frac{N^2}{R_g^{d+1}}=0 \quad\Rightarrow\quad R_g \sim N^{\frac{3}{d+2}}$$

So the exponent **$\nu_F = 3/(d+2)$.** Is it different from the random walk?

d=1	ν_F=1	EXACT!!!
d=2	ν_F=3/4	IT IS EXACT TOO!!! (complex analytical derivation)
d=3	ν_F=3/5=0.6	Numerics+Experiments show **ν=0.588**
d$\geq$4	ν_F=1/2	Random Walk!!!

These results show, in agreement with the simple geometrical arguments based on the fractal dimension of the intersection, that above d=4 self-intersections are negligible and the random walk result is obtained.
Below d=4 the Flory method, although approximate, performs amazingly well. How can we explain its quality?

Let’s consider P(R,N) as the “Boltzmann” factor associated with SAWs of length N and radius of gyration R. Then, their free energy is

$$\boldsymbol{F}(\boldsymbol{R},\boldsymbol{N}) = -\ln \boldsymbol{P}(\boldsymbol{R},\boldsymbol{N}) = \frac{\boldsymbol{R}^2}{\boldsymbol{N}} + \frac{\boldsymbol{N}^2}{\boldsymbol{R}^{\boldsymbol{d}}}$$

Maximizing the probability means minimizing the free energy,
Which is what has to be done to find the equilibrium state.

Then we can identify the two terms of the free energy as
an entropic and an energetic contributions:
the entropy is related to the random walk part, that has
no energetic characterization;
the energy is related to the volume exclusion.

The reason why Flory works so well is likely that both
energy and entropy are over-estimated!!!

Now that we have introduced the concept of free energy, we can go further and look better at the energies that are relevant for polymers.

We can imagine two kinds of energies: interactions between monomers and individual monomer interactions.

We deal first with single monomer energies.

The typical single particle energy is the chemical potential η: how much does it cost to add one monomer to the polymer?

The free energy of a polymer of length N can be written as

$$F(N+1) = (N+1)\eta - TS(N) = (N+1)\eta - k_B T \ln(\mathfrak{N}(N))$$

Where $\mathfrak{N}(N)$ is the number of conformations of a SAW of length N (which is made of N+1 monomers).

We can try to estimate $\mathfrak{N}(N)$ on a square lattice. Clearly, the SAW cannot co back on its trail at the first step, so that every step has 3 directions. This is an overestimate!!!
If we count only directed paths, that only go to the right and upward, then every step has two possible directions. This is an underestimate.

$$2^N \le \mathfrak{N}(N) \le 3^N$$

Both upper and lower bound grow exponentially with the length so the number of SAWs has to grow exponentially too. In particular it is

$$\mathfrak{N}(\boldsymbol{N}) = \boldsymbol{N}^{\gamma-1}\boldsymbol{\mu}^{\boldsymbol{N}}$$

Where γ is known as "entropic" exponent and is usually close to, but greater than, 1.

The constant μ is known as the connectivity constant of the SAW and it gives a measure of the average number of directions that every step can take (d=2, μ=2.68…).

Then we have

$$F(N+1)=(N+1)\eta-TS(N)=N\eta-k_BTN\ln\mu-k_BT(\gamma-1)\ln(N)$$

In the thermodynamic limit, $N\to\infty$, the free energy per particle is $f=\eta-k_BT\ln\mu$

What is more interesting is the associated partition function

$$Z(N+1)=z^{(N+1)}\mu^N N^{(\gamma-1)}$$

Because it allows the computation of the *grancanonical* partition function (z=fugacity=exp(-η/k$_B$T)<1; hence η>0)

$$\Xi(z)=\sum_{N=0}^{\infty}z^{(N+1)}\mu^N N^{(\gamma-1)}$$

Then we can compute the average number of monomers in the polymer

$$\langle N \rangle = z \frac{d}{dz} \ln \Xi(z) \propto (1 - z\mu)^{-1}$$

The number of monomers is finite as long as z<1/μ, and it diverges there: if the fugacity z=1/μ then we have a divergence.

Interpretation:

$$z = 1/\mu \quad \Rightarrow \quad \eta = k_B T \ln \mu$$

This means that, if the energetic cost of a monomer is perfectly compensated by the entropic gain, the length diverges.

We deal now with two-body (monomer-monomer) interactions.

First, we have to understand what kind of interactions we can expect to be relevant for polymers in solution.

There are two classes of interactions that are important: direct and effective interactions:

- direct interactions are well known, at least in principle: electrostatic interactions are a typical case and are of course extremely important. In particular they can be polar, dipolar or multipolar: the latter are usually neglected and are dealt with in terms of Van der Waals interactions.
- effective interactions are interactions that are “mediated” by the solvent; they are thermodynamic in nature.

Electrostatic interactions are usually easy to deal with:

the basis is the usual coulomb potential, possibly screened by the presence of ions in solutions to give an exponential cutoff to the interaction

$$\boldsymbol{V}(\boldsymbol{r}) = \frac{\boldsymbol{e}^{-\boldsymbol{r}/\boldsymbol{r}_0}}{4\pi\varepsilon\varepsilon_0 \boldsymbol{r}}$$

The Van der Waals interaction is simply treated by means of Lennard-Jones potentials between particles *i* and *j*

$$\boldsymbol{V}(\boldsymbol{r}_{ij}) = -\alpha\left[\left(\frac{\sigma}{\boldsymbol{r}_{ij}}\right)^6 - \frac{1}{2}\left(\frac{\sigma}{\boldsymbol{r}_{ij}}\right)^{12}\right]$$

where r_{ij} is the distance between the 2 particles

The Lennard-Jones potential is strongly repulsive close to contact distance (this mimics the hard-core repulsion of two solid objects); it is weakly attractive at large distances. The minimum of the potential is at $r_{ij}=\sigma$.

Dealing with effective interactions is instead trickier, because their formal derivation is difficult, and one has to resort to approximations that inevitably reduce the details of the effect. Only close inspection, and physical intuition, can help understand whether the approximations are legitimate and do not lead to the wrong results, or instead they are too strong.

Effective Interaction

The effective interaction of two particles is due to the fact that actually they interact with a third particle (or group of particles) but not with each other. Yet, if one observes the net effect, the result can be described by means of an interactions between the two particles. For example, particles A and B could have no mutual attraction, yet they could bind strongly to C. If we forget (in proper mathematical terms) about C, the net effect is that A and B are often close together and we can say that, at an EFFECTIVE level, they attract each other.

Formal derivation of effective interactions:

Let’s say that the degrees of freedom of the solvent are described by a collection of variables $\left\{\vec{r}_i\right\}$ and the solute particles are described by another collection of variables $\left\{\vec{R}_i\right\}$. These variables could be the positions but also the orientations and maybe other degrees of freedom.

The partition function is

$$Z = \int \prod_i d\vec{R}_i \int \prod_j d\vec{r}_j \exp\left(-\beta H\left[\left\{\vec{R}_i\right\},\left\{\vec{r}_j\right\}\right]\right)$$

Where $\boldsymbol{H}\left[\left\{\vec{\boldsymbol{R}}_i\right\},\left\{\vec{\boldsymbol{r}}_j\right\}\right]$ is the Hamiltonian (energy) of the system.

Looking for an effective energy means that we are looking for an expression of the kind

$$\boldsymbol{Z} = \int \prod_{i} d\vec{\boldsymbol{R}}_i \exp\left(-\beta \boldsymbol{H}_{eff}\left[\left\{\vec{\boldsymbol{R}}_i\right\}\right]\right)$$

Where we have

$$\boldsymbol{H}_{eff}\left[\left\{\vec{\boldsymbol{R}}_i\right\}\right] = -\boldsymbol{k}_B \boldsymbol{T} \ln\left[\int \prod_{j} d\vec{\boldsymbol{r}}_j \, \mathrm{e}^{-\beta \boldsymbol{H}\left[\left\{\vec{\boldsymbol{R}}_i\right\},\left\{\vec{\boldsymbol{r}}_j\right\}\right]}\right]$$

So we have an expression for the effective energy that depends on the temperature and that can not, rigorously, be written as the sum of two body interactions:

$$H_{eff}\left(\left\{\vec{R}_i\right\}\right) \neq \frac{1}{2}\sum_{i,j} V\left(\vec{R}_i, \vec{R}_j\right)$$

where $V\left(\vec{R}_i, \vec{R}_j\right)$ is the effective potential between two particles obtained for a system made of only two solute particles. Rather we must write

$$H_{eff}\left(\left\{\vec{R}_i\right\}\right) = \frac{1}{2}\sum_{i,j} V\left(\vec{R}_i, \vec{R}_j\right) + \left[H_{eff}\left(\left\{\vec{R}_i\right\}\right) - \frac{1}{2}\sum_{i,j} V\left(\vec{R}_i, \vec{R}_j\right)\right]$$

The term in square brackets contains intrinsically many-body interactions. If it is small with respect to the first term, made only of two body interactions, then we can approximate the energy by two-body interactions; else the situation remains more complex.

In the second case there is no rule, in general, to decide whether neglecting the term in square brackets is crucial to the physics of the system: maybe it changes the quantitative details but not the qualitative picture; maybe it changes also the important features of the system. A priori, there is no way to know.

Hydrophobicity is an effective interaction: simply stated, it is the aversion of some molecules with respect to water. The molecular origin of hydrophobicity are yet to be completely understood.

For our purposes it is important to remember that it is an effective interaction, that therefore it depends on temperature, pressure and other thermodynamic macroscopic parameters, and that in principle it cannot be reduced to a sum of two-body interactions (although the latest could be a fairly good approximation).

It is important to remember that hydrophobicity is very likely the most important interaction for living matter!!!

INTERACTION BETWEEN MONOMERS

To explore the effects of interactions between monomers we first have to analyze something simpler: the lattice gas.

The lattice gas represents a collection of particles free on a lattice, that interact when on nearest neighbor sites.

We assume that the attractive interaction is ε and that the chemical potential is η.

The Hamiltonian of the system is

$$\boldsymbol{H} = -\varepsilon \sum_{<i,j>} \boldsymbol{n}_i \boldsymbol{n}_j + \eta \sum_{i} \boldsymbol{n}_i$$

Where n_i is a variable that takes value 1 if site i is occupied by a particle, 0 otherwise.
As a first step we show that the lattice gas is equivalent to the Ising model, which allows us to understand the role of the chemical potential.

Let us make the following substitution: $\boldsymbol{n_i} = \frac{\boldsymbol{s_i} + 1}{2}$

The variable $\boldsymbol{s_i}$*=±1*, so it is a legitimate Ising variable. The Hamiltonian becomes

$$\boldsymbol{H} = -\frac{\varepsilon}{4} \sum_{<i,j>} s_i s_j - \left(\frac{z\varepsilon}{4} - \frac{\eta}{2} \right) \sum_i s_i$$

This is clearly the Hamiltonian of the Ising model in an external field

The Ising Hamiltonian is

$$H = -J \sum_{<i,j>} s_i s_j - h \sum_i s_i$$

The identification is now

$$J = \frac{\varepsilon}{4} \qquad h = \frac{z\varepsilon}{4} - \frac{\eta}{2}$$

Why is the chemical potential important?

We know that the Ising model is characterized by a low temperature phase that is ordered, and by a disordered high temperature phase. How does the crossover between the two take place?

We are usually taught that there is a phase transition at some critical temperature T_c such that the passage between the two is sharp. Actually, this is true if $h=0$. If instead $h\neq 0$ then the crossover is smooth and there is no rigorous phase transition.

It is clear therefore what is the role of the chemical potential: without it the lattice gas would correspond to an Ising model with external field: no phase transition.

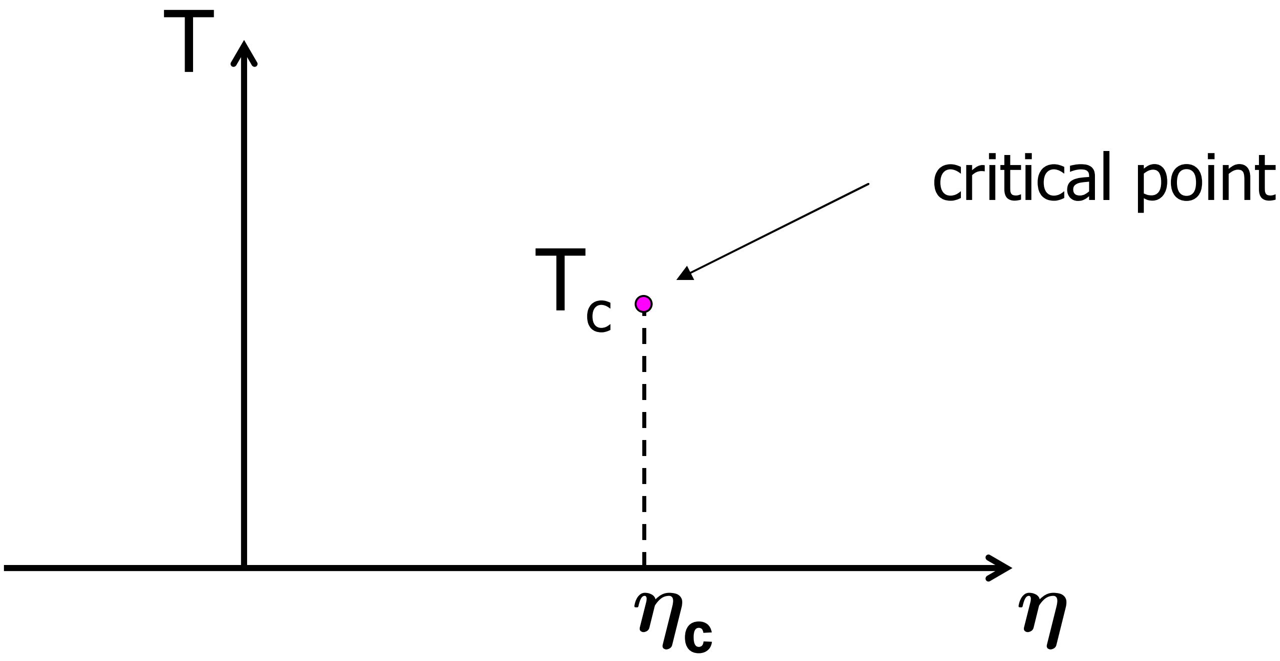
It is instead enough to set the chemical potential to

$$\eta = \frac{z\varepsilon}{2}$$

and the external field disappears. Hence we can have a phase transition between an ordered phase and a disordered one.

This is the second time where we find that the chemical potential is important to have criticality!!!

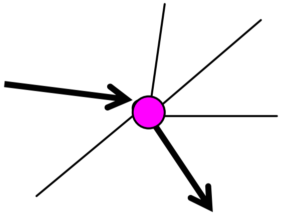
The global phase diagram is



The broken line represents a first order phase transition

We want to use the same (mean field) approach with a polymer. Where is the difference? The difference is that now monomers are tethered.

We define again the probability that a site is occupied as ρ, and that it is empty as 1- ρ. Yet we have to be careful: indeed an occupied site can be occupied in many different ways.



given an incoming direction, it has z-1 outgoing direction, in principle all equivalent to each other.

Therefore the probability of a site being occupied is

$$\rho = \sum_{i=1}^{z-1} \rho_i$$

where $\rho_\mathbf{i}$ is the probability to go in a given direction. Since all of them are equal we say that

$$\rho_i = \frac{\rho}{z-1}$$

The free energy per site is

$$f = u - Ts$$

where *u* is the internal energy and *s* the entropy.
The internal energy is the average of the Hamiltonian.
We have thererfore

$$u = -\varepsilon \frac{z-2}{2} \rho^2 + \eta\rho$$

where z-2 comes from the available number of nearest neighbors with which a site can interact (consecutive monomers do not interact!).

The entropy is instead

$$s = -k_B\left[\sum_{i=1}^{z-1}\rho_i \ln \rho_i + (1-\rho)\ln(1-\rho)\right] =$$

$$= -k_B\left[\sum_{i=1}^{z-1}\frac{\rho}{z-1}\ln\frac{\rho}{z-1} + (1-\rho)\ln(1-\rho)\right] =$$

$$= -k_B\left[\rho\ln\frac{\rho}{z-1} + (1-\rho)\ln(1-\rho)\right] =$$

$$= -k_B\left[\rho\ln\rho - \rho\ln(z-1) + (1-\rho)\ln(1-\rho)\right]$$

Overall the free energy becomes

$$\boldsymbol{f} = -\boldsymbol{\varepsilon}\frac{\boldsymbol{z}-2}{2}\rho^{2} + \left[\eta - \boldsymbol{k_B T}\ln(\boldsymbol{z}-2)\right]\rho +$$

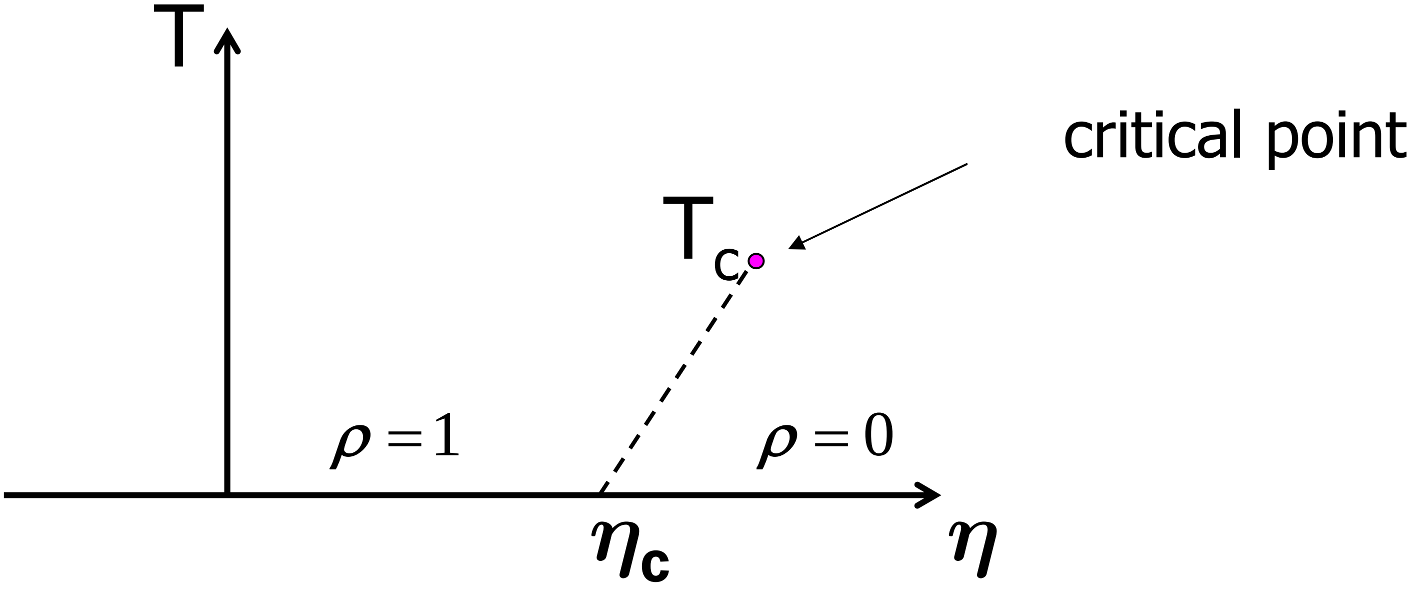
$$+\boldsymbol{k_B T}\left[\rho\ln\rho + (1-\rho)\ln(1-\rho)\right]$$

In the Ising language we have that the linear term becomes

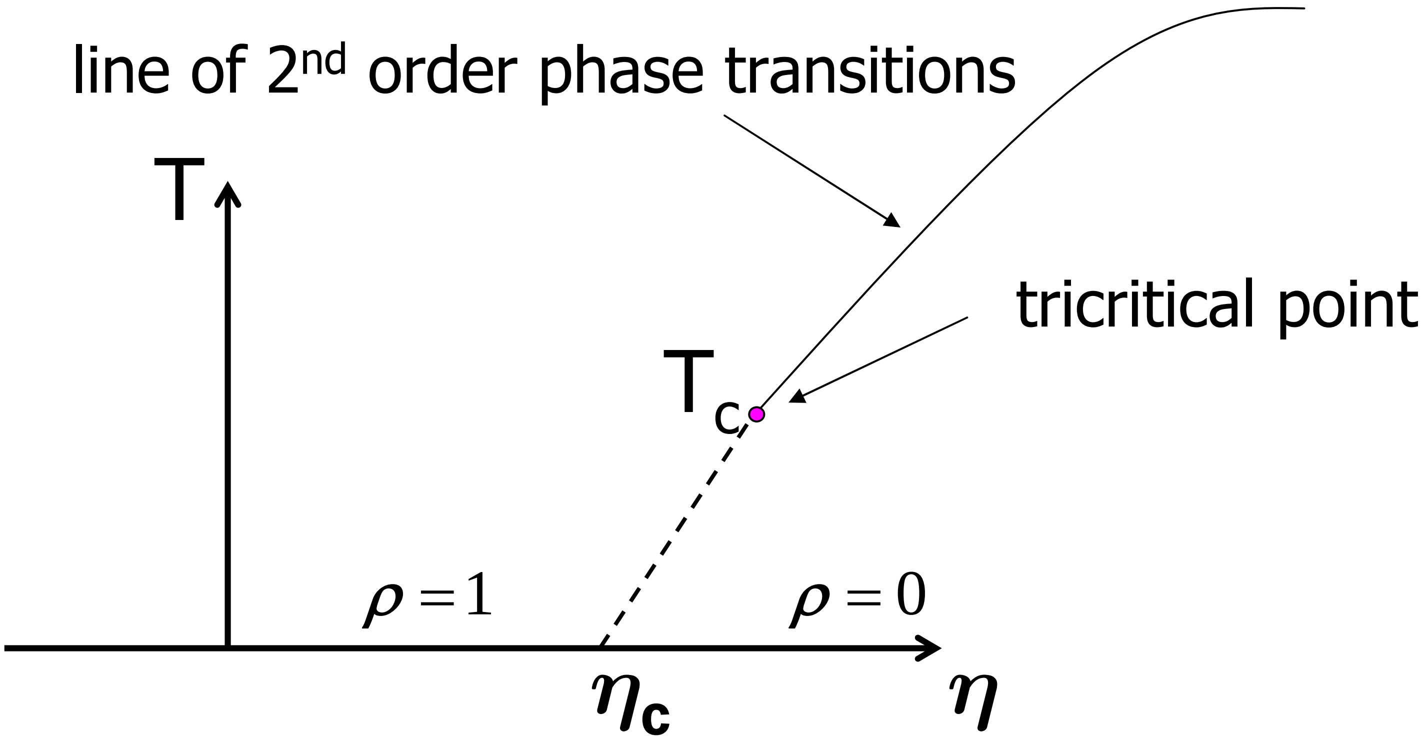
$$\left[\eta - \boldsymbol{k_B T}\ln(\boldsymbol{z}-2) - \frac{\boldsymbol{z\varepsilon}}{2}\right]\boldsymbol{m}$$

The linear term corresponds to an external field, and therefore we have a line of phase transitions if we set it to zero.
We obtain

$$\eta = \frac{z\varepsilon}{2} + k_B T \ln(z-2)$$



A more refined analysis tells instead that the phase diagram is like below:



Interpretation:

the low density region represents a region where there are just small polymers, of finite length (although enough to give a density different from zero);
the “critical” line represents a region where the length of the polymer diverges (remember the role of the chmical potential!!!);
the high density region represents a region where there is an infinite polymer that is very “dense” in the space: imagine a line that goes through most of the sites of the lattice!

Physically it means that, if we move over the critical line (the line of second order phase transitions), we have that the polymer is infinite and of moderate density. As we lower the temperature, at some point we meet the tricritical point, where a single state for the polymer is not admissible anymore: the system separates in two phases. The dense phase represents the polymer in a compact state, where it passes through most of the sites of the system; the dilute phase represents a phase where there are just strands of the polymer still in solution.

Essentially the polymer has collapsed from a swollen conformation, typical of SAWs, to a compact conformation.

A different interpretation, more easily related to experiments, goes as follows:

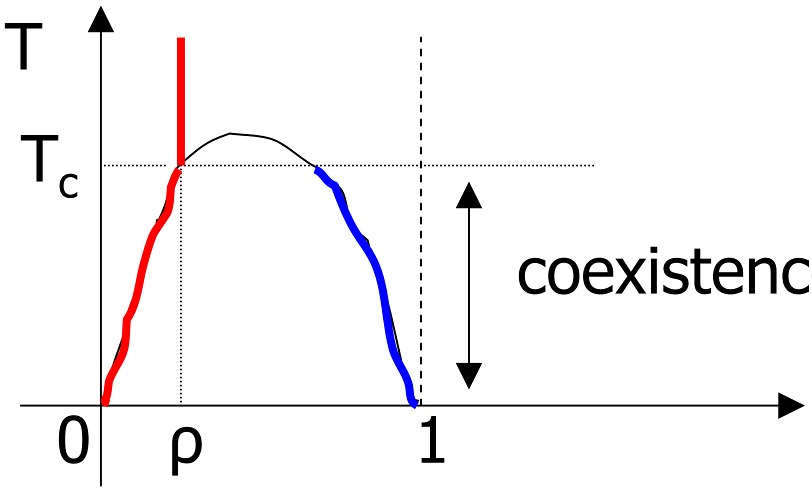
Let’s take a solution of polymers at a given density ρ, which is neither 0 nor 1. At very low temperatures the only two possible densities are (see phase diagram) only 1 and 0. Therefore the system cannot be in either… but in both!!!

Along the cut, it is possible to obtain any total density by splitting the system in different parts, each with a specific density, according to volume fractions

$$\rho_1 \frac{V_1}{V} + \rho_2 \frac{V_2}{V} = \rho$$

$$V_1 + V_2 = V$$

So, at low temperatures there is a part of the volume occupied by very dense polymers (compact polymers) and another part occupied by low-density polymers (swollen polymers). As temperature increases, the densities ρ_1 and ρ_2 change but as long as none of them matches ρ, the system is forced to be in a coexistence phase. When at last one of the two densities is equal to ρ, the system is again in a single phase.



coexistence in this region

Physics of Proteins

Proteins are polymers that, below a given “folding” temperature, take a compact conformation.

We have learnt in the previous parts that fundamental ingredients to have such a behavior are:
- self avoidance (so to correctly model polymers)
- attractive monomer-monomer interactions (so to allow the low temperature collapse).

Are we happy with these ingredients? We have to look at the phenomenology to decide whether we must add anything.

A collapsing polymer can go into one of the many possible compact conformations (akin to Hamiltonian walks). Instead, a folding protein goes into a single compact conformation, the native one.

Moreover, folding is a dynamical process: after synthesis, there is no temperature change to justify the folding. Simply, at body temperature, proteins are synthesized in a non-native conformation and they have to dynamically change it until they reach the native state. Since proteins are small and in solution, these changes are ruled by brownian motion.

Levinthal paradox: even if every conformational change could take place in a femtosecond, the time to fold a protein would be larger than the age of the Universe… a little bit too much.

Anfinsen solution to the paradox: folding is Brownian, but in a potential landscape where the native state is at the bottom of a very large basin of attraction. This is the funnel concept. Such an energy funnel is defined by energies encoded in the amino-acid sequence. Anfinsen showed this by folding proteins in vitro: the only ingredient is the sequence (there are no special “folding machines” that help proteins in vitro).

Dogma: the sequence dictates the native structure!

Anfinsen won the Nobel prize for this discovery.

The missing ingredient: the sequence!!!

There are 20 naturally occurring amino-acids (synthesized by living organisms).

Roughly half of them are hydrophobic and attract in water. They provide the driving force for the first stages of the collapse (just as for homopolymers).
The other half are polar, and they love water (that is a polar molecules).

It’s a subtle mix of the two kinds, in particular orders, that provide the missing ingredients.

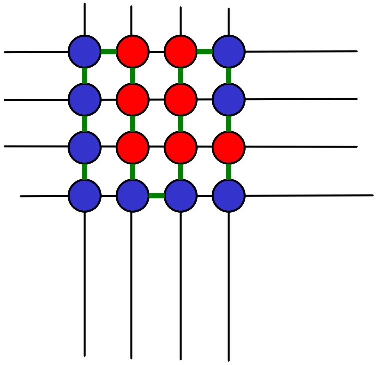
The “Protein Folding Problem” consists in finding the right key to decode the sequence.
There are different strategies:

1. All atom brute force simulations: let’s take the protein with all of its atoms (and maybe the solvent too) and let’s solve Newton’s equations. It can (and likely it will) work, but:
   a) do we really learn the code?
   b) computers are not yet powerful enough
2. Empirical methods: is two proteins have a similar sequence and we know the native state of one of them, we may guess the native state of the other.

Maybe we can try to learn something simpler by simple models. Important: as we introduce simplifications, we can ask only simple questions!

First simple model: the HP model.

As stated, we define only two kinds of amino-acids: hydrophobic and polar (H and P). We add the ingredient that H amino-acids attract each other (it’s an effective interaction!). Then proteins are modeled on a lattice (two or three dimensions, it does not matter at a first stage).



🔵 P

🔴 H

In this conformation there are, for example, 4 HH contacts: the energy is –4.

It has been shown that, out of the 2^N possible sequences (N is the protein length), just a few are “good” sequences: they have a unique native state (the ground state for this model). Therefore only a small number of possible sequences can work as proteins. This is not yet the “decoding” of the sequence, because looking at them we still do not understand.

A further important result is that of all the possible compact conformations, only a few are chosen by the “good” sequences as native states. This is important because the concept of “fold family” is an important one. It is seen that of many proteins whose structure is known, many have very similar structure even if the sequence is quite different.

Moreover, the HP model predicts proteins with a hydrophobic core (as seen experimentally) that fold below a certain temperature.

Unfortunately the HP model is not the best to address the dynamics: it is on a lattice, so that a natural dynamics cannot be defined.

The same results are found in 2 and 3 dimensions, and even by enlarging the amino-acid alphabet to 3 or 5 letters (details change but not the main message).

What if we want to look to more refined models, or to look at real proteins, without using all the atomistic information?

We have to resort to simplified off-lattice models.

Since we do not want all the details, where do we introduce simplifications?

We simplify both the degrees of freedom and the energies.

1. every amino acid is treated as a spherical ball. A protein is then a chain of beads, each one with a label (its amino-acid species).
   The relevant degrees of freedom are the two rotation angles φ and θ that characterize the direction of every protein bond with respect to the previous one.

2. The energies are of two kind: local and non-local. Local energies depend on the local bending angles, E(φ,θ). Non-local energies are related to contacts between amino-acids that are not consecutive along the chain.

Determination of the energy functions:
although complex, E(φ,θ) can be determined by the physics of the peptide bond (the link between consecutive amino-acids).

The non-local interaction is instead more difficult to obtain: it could be derived from first principles, but that would be still too difficult. So we need some different way to determine it.
The spatial form is the usual Lennard-Jones potential

$$\boldsymbol{V}_{\alpha,\beta}(\boldsymbol{r}) = \boldsymbol{\varepsilon}_{\alpha,\beta}\left[2\left(\frac{\boldsymbol{r}_0}{\boldsymbol{r}}\right)^{12} - \left(\frac{\boldsymbol{r}_0}{\boldsymbol{r}}\right)^{6}\right] \qquad \textbf{\textit{if attractive}}$$

$$\boldsymbol{V}_{\alpha,\beta}(\boldsymbol{r}) = 2\boldsymbol{\varepsilon}_{\alpha,\beta}\left(\frac{\boldsymbol{r}_0}{\boldsymbol{r}}\right)^{12} \qquad \textbf{\textit{if repulsive}}$$

The values of $\varepsilon_{\alpha,\beta}$ depend on the two amino-acid species under consideration, α and β. How can these values be determined?

Idea: we can learn them on known native states, 10000 (rough numbers) of which are known by X-ray crystallography, NMR and other techniques.

We look at, say, 1000 native states and see that in most cases species α and β are closer to each other than a given threshold distance (a parameter of the procedure). Then we can say that they have an attractive interaction and $\varepsilon_{\alpha,\beta}$ <0.
If they are most often above that threshold then $\varepsilon_{\alpha,\beta}$ <0.
If they are equally distributed above and below, their interaction is negligible.

Of course the real “learning” techniques are much more complex, technically, but their essence is what stated in the previous slide.

Do these techniques work? Yes and no.

If we try to fold a protein using these interactions, we will be successful with a score of 60-70% on proteins that were part of the learning set (the potentials are indeed optimized for those proteins), but we fail most of the time as we try to apply them to proteins not in the training set. Consequently, we cannot trust them much for proteins whose native structure is not even known.

Although all these results show that our understanding of proteins is still limited, and our ability to use simple models to describe them is not what theoretical physicists are used to, this is a consequence of the relatively young age of the domain.

Levy probability distributions

The Central Limit Theorem holds for indipendent identically distributed (iid) random variables with finite first and second moments

So, we can break the CLT if one (or both) of the above conditions is broken.

Let us take $$X = \sum_{i=1}^{N} x_i$$

$$\langle x \rangle = x_0$$

$$\langle x^2 \rangle = \sigma^2 + x_0^2$$

where σ is the variance of the distribution.
For a symmetric distribution ***x₀=0***.

With *x_i* taken all from a given probability distribution *p(x),* with first and second moments

$$\langle x \rangle = x_0$$

$$\langle x^2 \rangle = \sigma^2 + x_0^2$$

The Central Limit Theorem (for physicists!)

$$X = \sum_{i=1}^{N} x_i$$

With x_i taken all from a given probability distribution *p(x)*. Then $P_N(x)$ is given by

$$P_N(X) = \int dx_1 \int dx_2 \ldots \int dx_N \prod_{i=1}^{N} p(x_i)\delta(X - \sum_{j=1}^{N} x_j)$$

The Central Limit Theorem (for physicists!)

$$X = \sum_{i=1}^{N} x_i$$

With x_i taken all from the probability distribution *p(x)*.
Then the second moment of *P(X)* is

$$\left\langle X^2 \right\rangle = \left\langle \sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j \right\rangle = \sum_{i=1}^{N} \left\langle x_i^2 \right\rangle + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left\langle x_i x_j \right\rangle =$$

$$= N\sigma^2 + N^2 \left\langle x \right\rangle^2 = N\sigma^2$$

By taking the Fourier transform of both hands of the equation we obtain

$$\tilde{P}_N(q) = \left[\tilde{p}(q)\right]^N$$

And the antitransform, in the limit of large *N*, is

Physique Statistique de
Biomacromolecules