Conformational Space, Energy Minimization and Molecular Mechanics

Conformational space

1. Definitions
2. Rubik's snake
3. Proteins
4. Levinthal's paradox

1. Definitions

Conformation is the set of 3N coordinates of all N atoms.

Conformational space is the space of all possible conformations

Questions

How big is conformational space?

Can proteins sample all conformational space?

Can proteins fold by random search?

If not, how do they fold?

2. Rubik's snake

Consider Rubik's snake as a model of protein:

   It consists of 24 fragments

   Folded conformation has a function of "rolling"

   All other conformations are not functional

Conformational space:

   Out of $4^{23} \approx 7 \cdot 10^{13}$ conformations,

   $\approx 10^{13}$ are spatially possible.

One sec per conformation means $\approx$ 400 000 yrs for all possible conformations

3. Proteins

Protein:

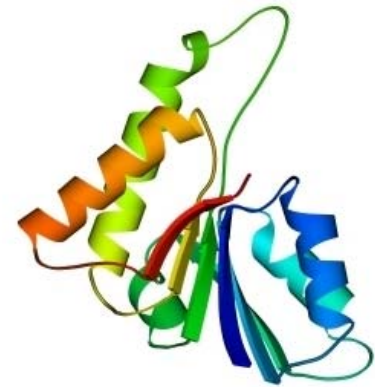   One folded functional conformation

      Many unfolded non-functional

Conformational space for 100 aa protein:

      $\approx 10^{100}$ conformations

One ps per conformation gives $\approx 10^{80}$ years

Proteins apparently do not fold by sampling all conformations

3. Proteins

Experimental facts:

Proteins fold into their native structure from microseconds to hours

Proteins fold into their native structure during biosynthesis, renaturation, after chemical synthesis. Then, native structure seems to be the energy global minimum

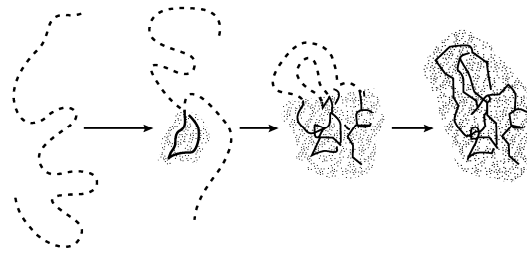There is a free-energy barrier between native and unfolded structures
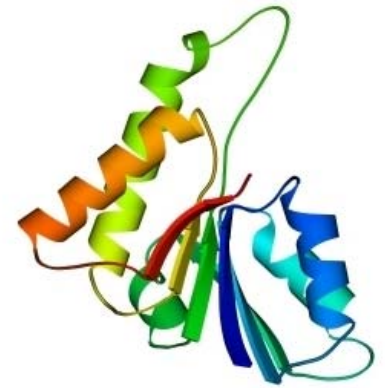
4. Levinthal's paradox

If native structure is the energy global minimum, then how proteins can find it if they don't have time to sample all conformations?

4. Levinthal's paradox

Solution:

Nucleation mechanism
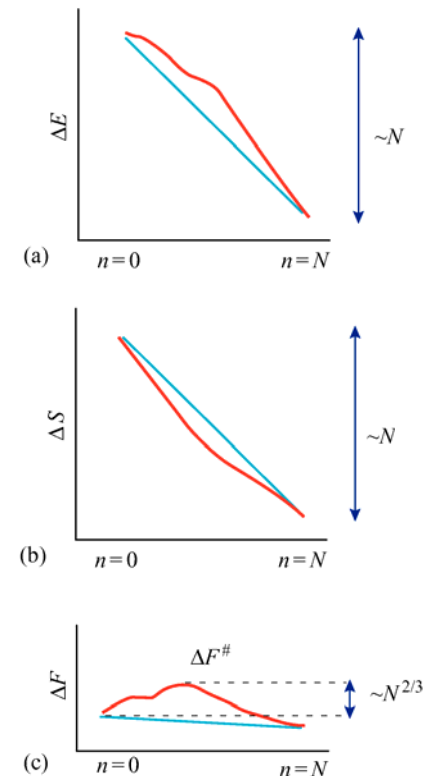
$$\Delta E_n \sim -n + \mu_n n^{2/3}$$

$$\mathsf{T}\Delta S_n \sim -n - \lambda_n n^{2/3}$$

$$\Delta F_n = \Delta E_n - T\Delta S_n$$

$$\Delta F^\# = \max\{\Delta F_n\} \sim N^{2/3}$$

$$time = 10ns \cdot exp\{(0.5 - 1.5)N^{2/3}\} \ll 10ns \cdot 10^N$$

Thus, we automatically have fast folding pathway to global minimum. Additional pathways can only speed up the folding

Mathematics

1. Taylor approach in 1D
2. Functions in R^N
3. The gradient
4. Taylor approach of f: R^N ->R
5. Energy as function E: R^N ->R
6. Newton-Raphson method
    1. Steepest descent
    2. Conjugate gradient

1. Taylor approach in 1D

Given f, n+1 times differentiable, and the polynomial P(x), as:

$$P(x) = f(a) + f'(a)(x-a) + \frac{1}{2} f''(a)(x-a)^2 + \ldots + \frac{1}{n!} f^{(n)}(a)(x-a)^n$$

There is a function h(x), such that:

$$f(x) = P(x) + h(x)(x-a)^n$$

$$\lim_{x \to a} h(x) = 0$$

Proof

Use the following theorem: there is z such that g'(z)(b-a)=g(b)-g(a)
By differentiating up to n and defining g=f$^{(n)}$, there is a z, such that the following h
function fulfills the theorem

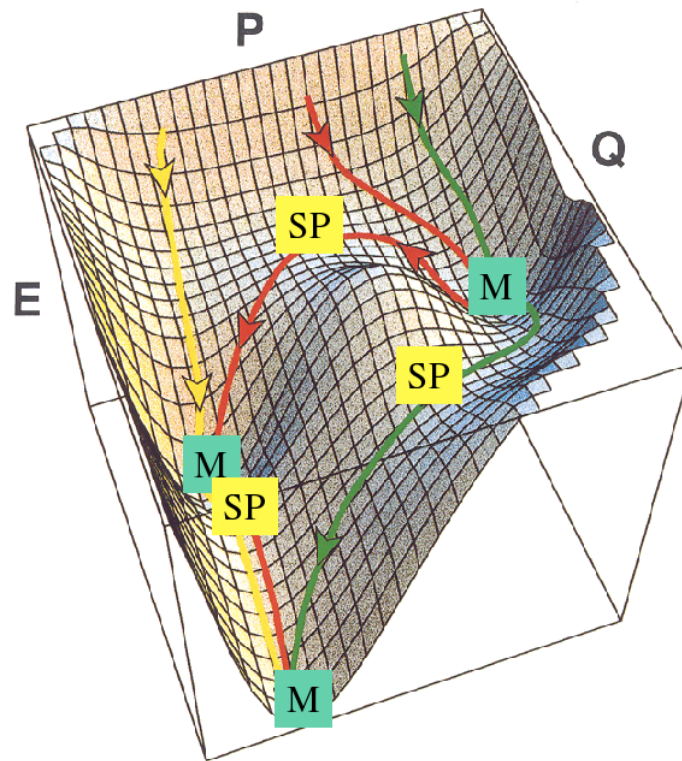$$h(x) = \frac{f^{(n+1)}(z)}{(n+1)!}(x-a)$$

2. Functions in R^N

$$f : R^N \to R$$

$$D_i f = \frac{\partial f}{\partial x_i} = \lim_{dx_i \to 0} \frac{f(x_1,..(x_i + dx_i),..x_N) - f(x_1,..x_i,..x_N)}{dx_i}$$

Example

3. Gradient

Given f:R^N->R, we say f is differentiable in a if for all v in R^N there is a linear application T$_v$:R^N->R, such that

$$\lim_{h \to 0} \frac{f(a + hv) - f(a)}{h} = T_a(v)$$

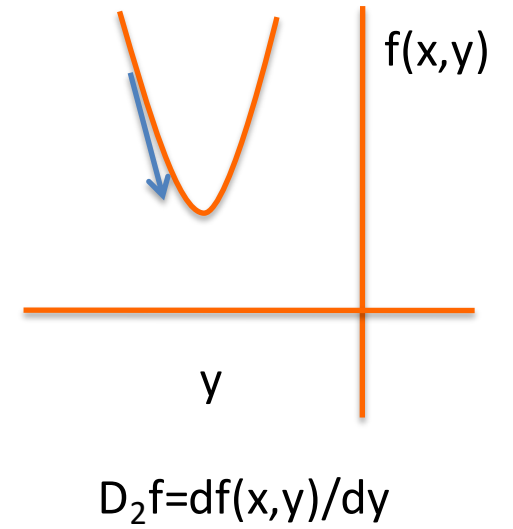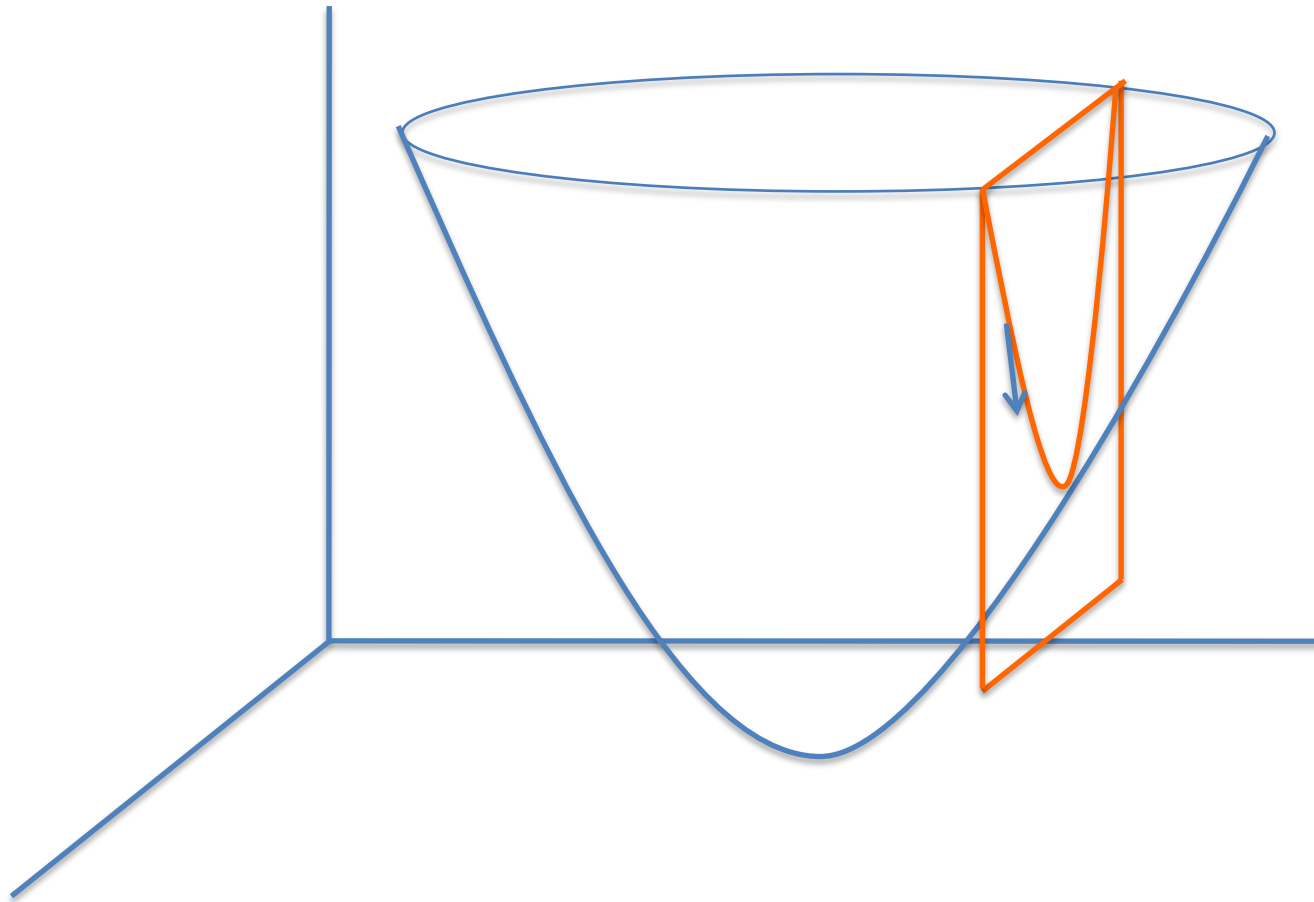It is easy to see that:

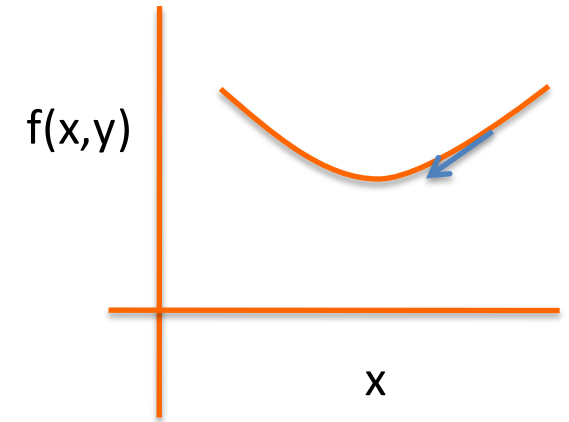$$\lim_{h \to 0} \frac{f(a + he_i) - f(a)}{h} = T_a(e_i) = D_i f(a)$$

Then we end with the definition of gradient

$$T_a(v) = T_a\left(\sum v_i e_i\right) = \sum v_i T_a(e_i) = \sum v_i D_i f(a) = \left(D_1 f(a), ... D_N f(a)\right)(v_1, ... v_N)$$

$$\left(D_1 f(a), ... D_N f(a)\right)(v_1, ... v_N) = \nabla_a f * v$$

3. Gradient

$$D_2 f = df(x,y)/dy$$

3. Gradient

$f(x,y)$

x
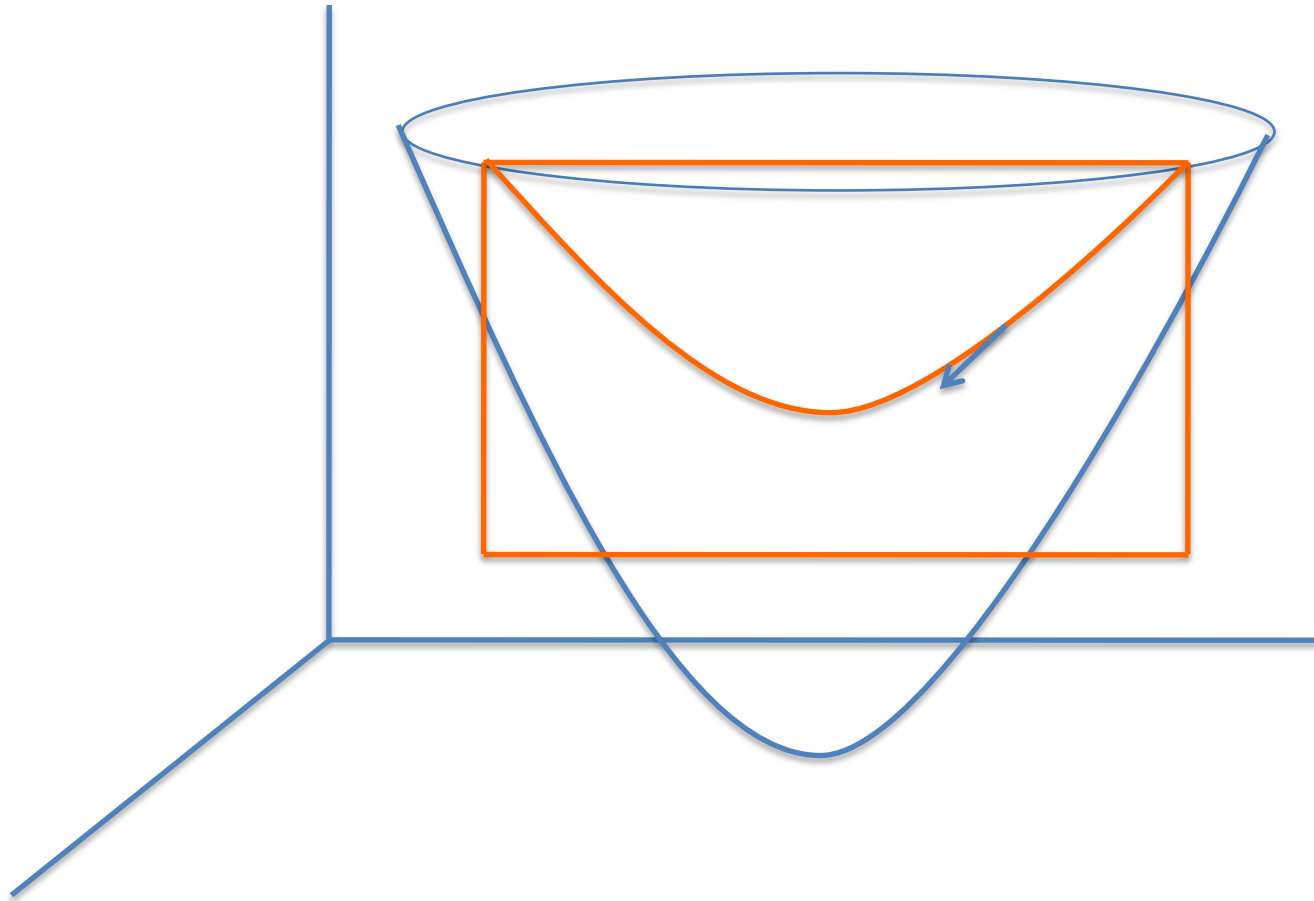
$D_1f=df(x,y)/dx$

3. Gradient

$$f = cte \Rightarrow df = 0$$

$$df = (\nabla f) * (\pm dv) \Rightarrow (\nabla f) \perp (\pm dv)$$

4. Taylor approach of f: R^N ->R

Chain Rule:

Let be g:R->R^N a curve in R^N, and f a function differentiable of R^N in R.
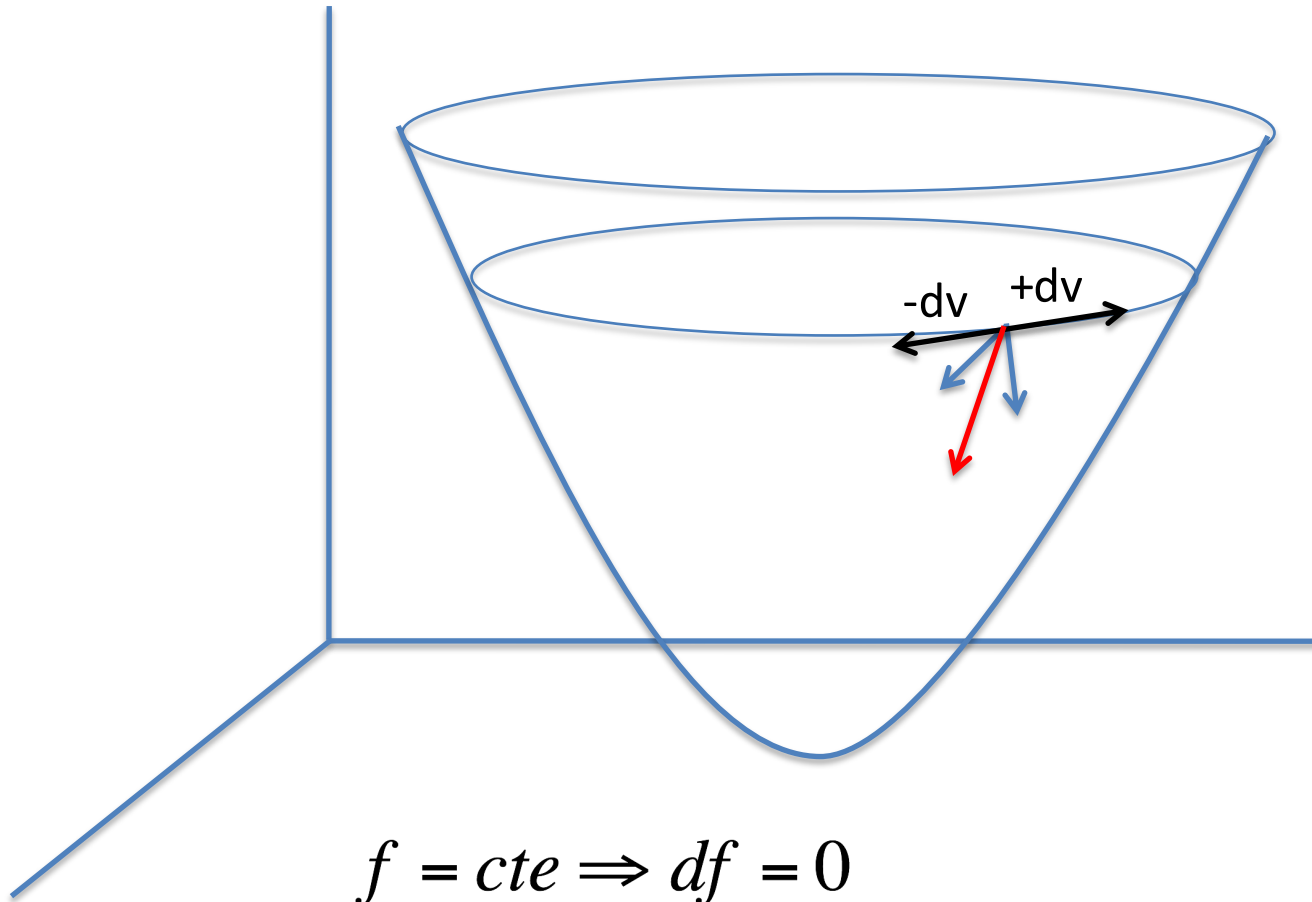We define the composition f*g:R->R by f*g(x)=f(g(x))
Then (f*g)´ is:

$$\left(f \circ g\right)' = \lim_{h \to 0} \frac{f(g(x+h)) - f(g(x))}{h} = \lim_{h \to 0} \frac{f(g(x) + hg'(x)) - f(g(x))}{h} = \nabla_{g(x)} f * g'$$

Then, let be b and a in R^N. We define the curve g(t):R-> R^N as g(t)=a+t(b-a) and
the function F as F=f*g. Therefore, F(0)=f(a), F(1)=f(b), and by Taylor approach:

$$F(1) = F(0) + F'(0) + \frac{1}{2}F''(0) + \dots \frac{1}{n!}F^{(n)}(0)$$

$$f(b) = f(a) + (\nabla f)(b - a) + \frac{1}{2}\nabla(\nabla f * (b - a)) * (b - a) + \dots$$

$$f(b) = f(a) + \sum_i (D_i f(a))(b_i - a_i) + \frac{1}{2}\sum_j\sum_i (D_{ji} f(a))(b_i - a_i)(b_j - a_j) + \dots$$

$$f(b) = f(a) + \sum_i (D_i f(a))(b_i - a_i) + \frac{1}{2}(b - a)H(b - a)^t + \dots$$

H is named Hessian matrix

5. Energy function

$$V = \frac{1}{2}\sum_i K_i^{bond}\left(d_i - d_{0,i}\right)^2 + \frac{1}{2}\sum_i K_i^{angle}\left(\alpha_i - \alpha_{0,i}\right)^2 + \frac{1}{2}\sum_i K_i^{dihedral}\left(\omega_i - \omega_{0,i}\right)^2$$

$$+ \frac{1}{2}\sum_i K_i^{torsion}\cos(\lambda_i\phi_i + \delta_i) + \frac{1}{4\pi\varepsilon}\sum_i\sum_{j>i}\frac{q_i q_j}{r_{ij}} + \sum_i\sum_{j>i}\left(\frac{C_6(i,j)}{r_{ij}^6} - \frac{C_{12}(i,j)}{r_{ij}^{12}}\right)$$

$$+ \sum_i\left(Hydrogen - Bonds\right) + \sum_i\left(\pi - \pi\right)$$

The quadratic terms show the potential origin in non-correlated internal coordinates, being their constants derived from the second derivative of the potential

6. Newton-Raphson approach

Given a function f: R->R, there is an iterative approach to find the zeros (x such as f(x)=0), defined by the series:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Proof

By Taylor's expansion up to 2on order:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + h(x)(x - x_0)$$

$$f(x_0) + f'(x_0)(x - x_0) \rightarrow 0$$

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}$$

6. Newton-Raphson approach

The same approach is valid for obtaining the zeros of the derivative, but the series is then:

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$$

When the function is in R^N the series is derived from Taylor's expansion for functions in R^N

$$x = x_0 - f'(x_0)\left(f''(x_0)\right)^{-1}$$

$$x = x_0 - \left[H^{-1}\right]\sum_i \left(D_i f(x_0)\right)$$

The method is modified with a lambda to perform small steps and ensure convergence

$$x_{n+1} = x_n - \lambda\left[H^{-1}\right]\sum_i \left(D_i f(x_n)\right)$$

6.1. Steepest Descent

Lambda is selected and modified at each step to fall quickly towards the minimum, by increasing when $x_{n+1} < x_n$ or decreasing and changing sign when $x_{n+1} > x_n$

$$x_{n+1} = x_n - \lambda \nabla f(x_n)$$

6.2. Conjugate gradient

Additional modification of steepest descent implies to store several gradients and modify the new direction with the previous

$$x_{n+1} = x_n - \lambda \nabla f(x_n) - \sum_i \lambda_i \nabla f(x_{n-i})$$

Physics

1. Molecular mechanics
2. Phase space
3. Maxwell-Boltzmann distribution of velocities
4. Molecular dynamics
    1. Verlets' algorithm
    2. Ergodic hypothesis
    3. Periodic boundary conditions
5. Monte-Carlo simulation
6. Simulation annealing

1. Molecular Mechanics

$$V = \frac{1}{2}\sum_i K_i^{bond}\left(d_i - d_{0,i}\right)^2 + \frac{1}{2}\sum_i K_i^{angle}\left(\alpha_i - \alpha_{0,i}\right)^2 + \frac{1}{2}\sum_i K_i^{dihedral}\left(\omega_i - \omega_{0,i}\right)^2$$

$$+ \frac{1}{2}\sum_i K_i^{torsion}\cos(\lambda_i\phi_i + \delta_i) + \frac{1}{4\pi\varepsilon}\sum_i\sum_{j>i}\frac{q_iq_j}{r_{ij}} + \sum_i\sum_{j>i}\left(\frac{C_6(i,j)}{r_{ij}^6} - \frac{C_{12}(i,j)}{r_{ij}^{12}}\right)$$

$$+ \sum_i\left(Hydrogen - Bonds\right) + \sum_i\left(\pi - \pi\right)$$

1. Molecular Mechanics

$$P_i = \frac{e^{-\beta E_i}}{Z} \qquad \text{and} \quad S = -R\sum_i P_i \times \ln(P_i)$$

where

$$\beta = \frac{1}{kT} \quad ; \quad Z = \sum_{i=1}^{\infty} e^{-\beta E_i} \quad ; \quad k = R/Na$$

$$E_i = K_i + V_i = \sum_j \frac{p_j^2(i)}{2m_j} + V_i(r_1, r_2, \dots r_n)$$

1. Molecular Mechanics

$$\langle A \rangle = \sum_{i \in \mu-states} A_i P_i = \int A(\vec{p},\vec{r}) \rho(\vec{p},\vec{r}) \partial p \, \partial r$$

$$\langle A \rangle = \int A(\vec{p},\vec{r}) \frac{e^{-E(\vec{p},\vec{r})/k_B T}}{Z} \partial p \, \partial r$$

2. Phase space/Conformational space

PHASE SPACE is the space of all possible states, defined by 6N coordinates (3N positions and 3N momenta). Neglecting the 3N moment, the space defined by the 3N positions is the CONFORMATIONAL SPACE

$$\vec{r} = r^N = (x_1, y_1, z_1, \ldots x_N, y_N, z_N)$$

$$\vec{p} = p^N = (p_{1x}, p_{1y}, p_{1z}, \ldots p_{Nx}, p_{Ny}, p_{Nz})$$

3. Maxwell-Boltzmann distribution

For an ideal gas E=K and V=0, then

$$1 = \iiint_{\infty} \frac{e^{-E(\vec{p},\vec{r})/k_B T}}{Z} \partial p\, \partial r = \iiint_{\infty} \frac{e^{-\sum_j p_j^2 / 2m_j k_B T}}{Z} \partial p$$

$$1 = \frac{1}{Z}\prod_i \iiint_{\infty} e^{-\left(p_{ix}^2 + p_{iy}^2 + p_{iz}^2\right)/2m_i k_B T} dp_{ix} dp_{iy} dp_{iz}$$
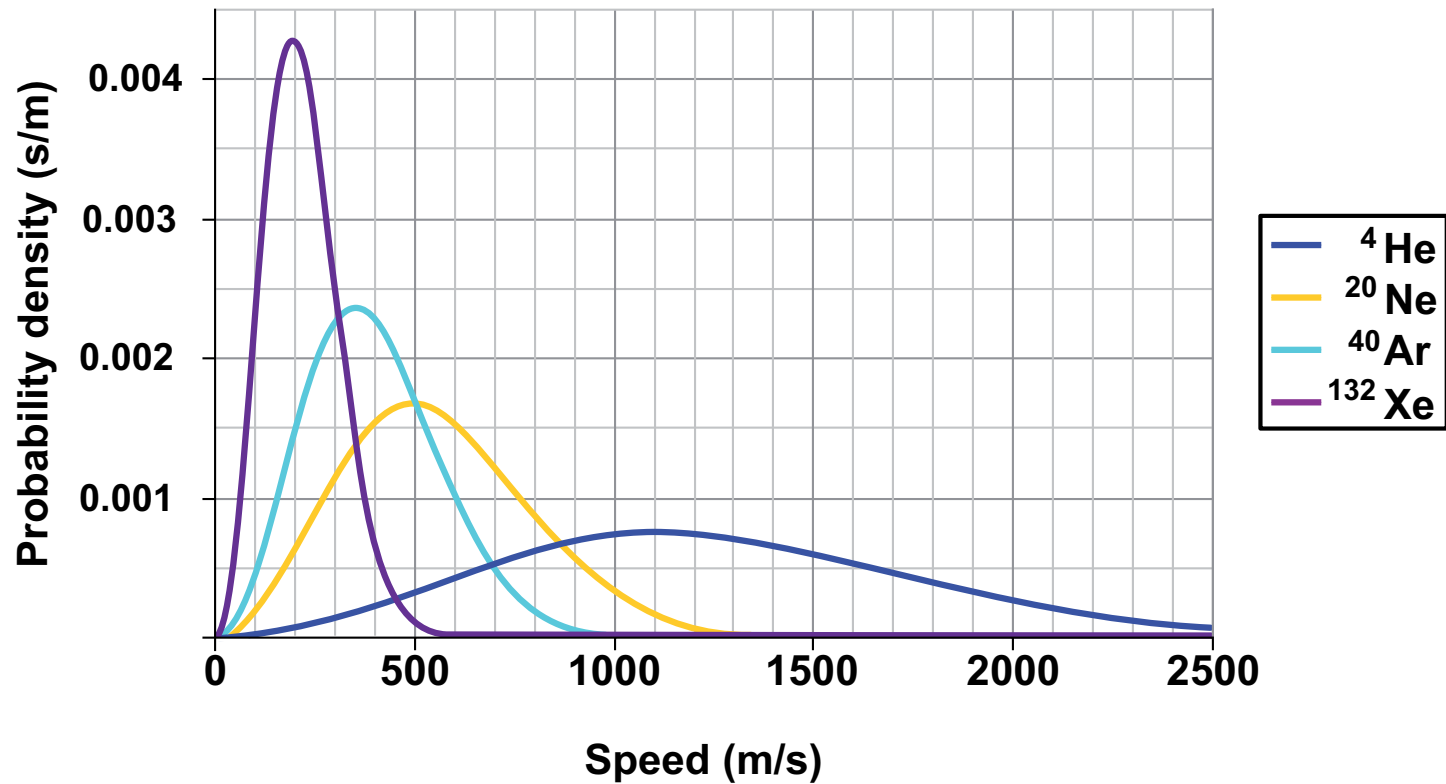
$$Z = \prod_i \left(\frac{1}{2\pi m_i k_B T}\right)^{3/2}$$

$$f_i(px, py, pz) = \left(\frac{e^{-p_{ix}^2/2m_i k_B T}}{\sqrt{2\pi m_i k_B T}}\right)\left(\frac{e^{-p_{iy}^2/2m_i k_B T}}{\sqrt{2\pi m_i k_B T}}\right)\left(\frac{e^{-p_{iz}^2/2m_i k_B T}}{\sqrt{2\pi m_i k_B T}}\right)$$

This is a normal distribution with variance $mk_B T$ and $m_i=m$ for all i

3. Maxwell-Boltzmann distribution

Normal distribution around 0, with variance mk_BT and $m_i=m$ for all i

Maxwell-Boltzmann Molecular Speed Distribution for Noble Gases

4. Molecular Dynamics

$$a = \frac{dv}{dt}$$

$$v = \int a\,dt = at + v_0$$

$$a = \frac{F}{m} = \frac{1}{m}\left(\frac{\partial V}{\partial x}\right)$$

$$v = \frac{dx}{dt}$$

$$x = \int (at + v_0)\,dt = \frac{1}{2}at^2 + v_0 t + x_0 = \frac{1}{2m}\left(\frac{\partial V}{\partial x}\right)t^2 + v_0 t + x_0$$

$$\vec{r} = \frac{1}{2m}(\nabla V)t^2 + \vec{v}_0 t + \vec{r}_0$$

4.1. Verlet's algorithm

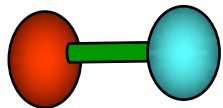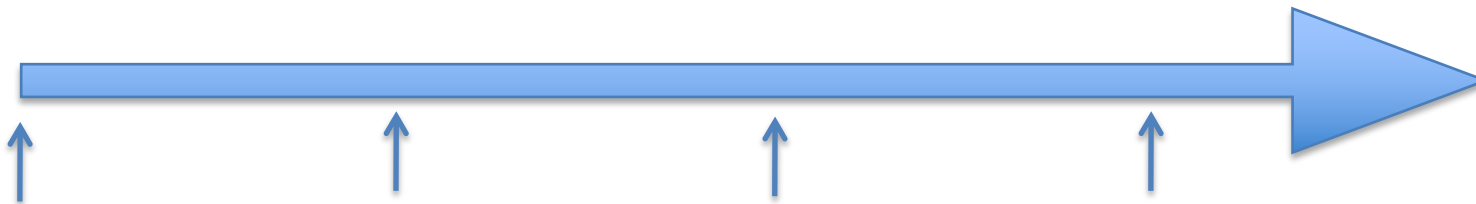$$a = \frac{F}{m} = \frac{1}{m}(\nabla F)$$

$$r(t + dt) = r(t) + v(t)dt$$

$$v(t + dt) = v(t) + adt = v(t) + \frac{dt}{m}(\nabla F)$$

$$r(t + 2dt) = r(t + dt) + v(t + dt)dt$$

.....

r_0 are the X-ray coordinates
v_0 are selected with a random function using a Maxwell-Boltzmann distribution

$$F = -k(d - d_0)$$

$$E = \frac{1}{2}k(d - d_0)^2$$

Vibration determines dt
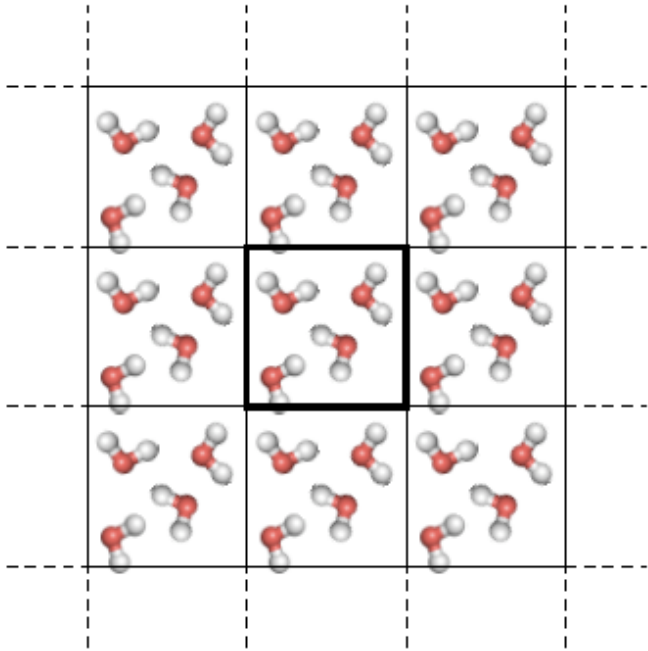For dihedral of rotation dt is around 2 fs

4.2. Ergodic hypothesis

$$\langle A \rangle = \sum_{i \in \mu-states} A_i P_i = \int A(\vec{p},\vec{r})\rho(\vec{p},\vec{r})\partial p\,\partial r$$

$$\overline{A} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^{\tau} A(\vec{p}(t),\vec{r}(t))dt$$

$$\langle A \rangle = \overline{A}$$

4.3. Periodic Boundary conditions

Limited distance of interactions:
1 atom cannot interact with itself (a)
1 molecule cannot interact with itself (b)

Cut-off determined by long distance interactions
Coulomb energy is the longest
Cut-off around 12-15 Angstroms
Minimum box-edge L/2 > 15A (a) or L>D+30A (b) with
D the maximum diameter of the protein

5. Monte Carlo simulations

•Start at original coordiantes r_0 and calculate the energy (E)
•Select few variables (i.e. distances, angles, dihedrals) and or atom coordinates (i.e. water or ion atoms) and change them with a small deviation.
•Calculate the new energy (E')
•Apply a criterion of acceptance with probability P using a random function:

$$P = \begin{cases} 1 & \text{if } E' \leq E \\ e^{-\Delta E/k_B T} & \text{if } E' > E \end{cases}$$

* Note: by increasing the temperature, large changes of conformation increase the probability to be accepted

6. Simulated Annealing

Features:
- E=K+V is constant
- $K=3/2\ Nk_BT$, with N= number of atoms
- If T=0, E=V
- When externally forcing T to be very high at a given conformation, we are increasing E
- For a given E, searching the conformation with maximum velocities (T) we obtain the minimum potential V
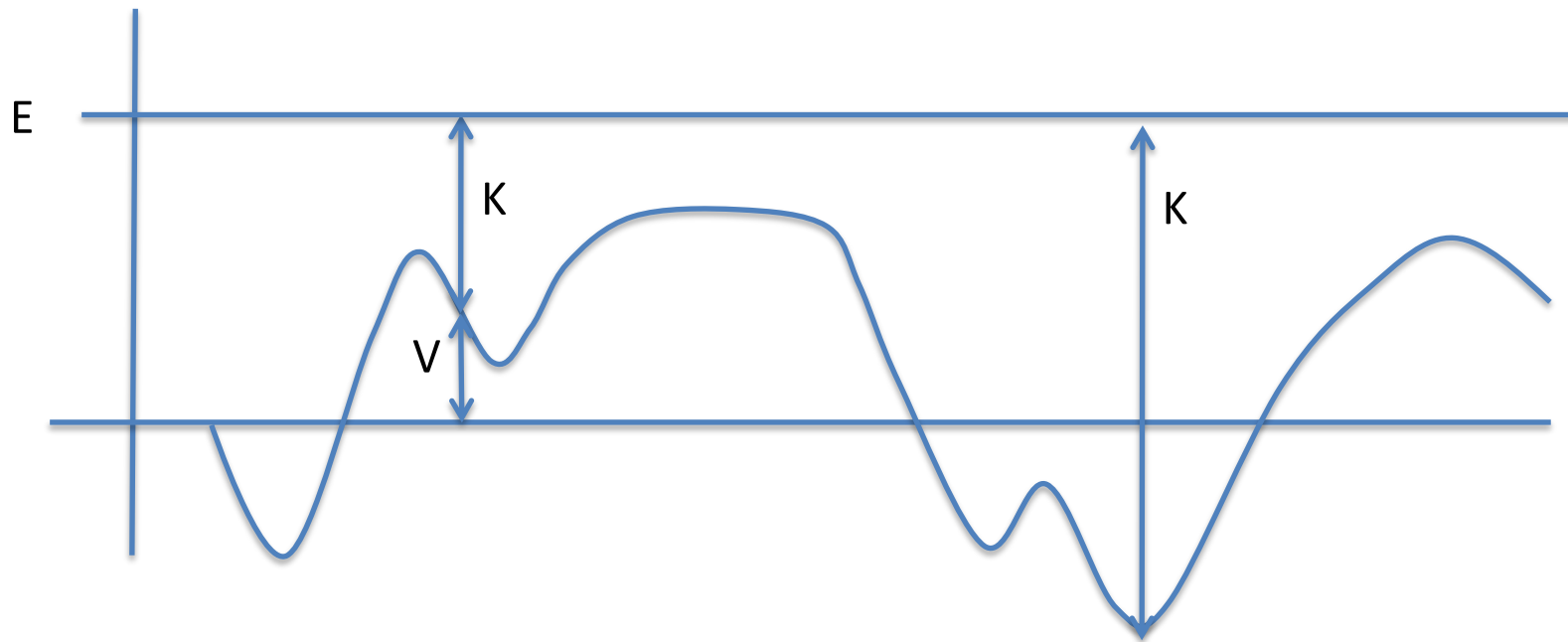
6. Simulated Annealing

Method:
- Start at T=0 and increase to T=1000 K
- Run short MD or MC simulation
- Select conformers with maximum K
- Run Steepest Descent
- Increase T to 1000K and repeat iteration.
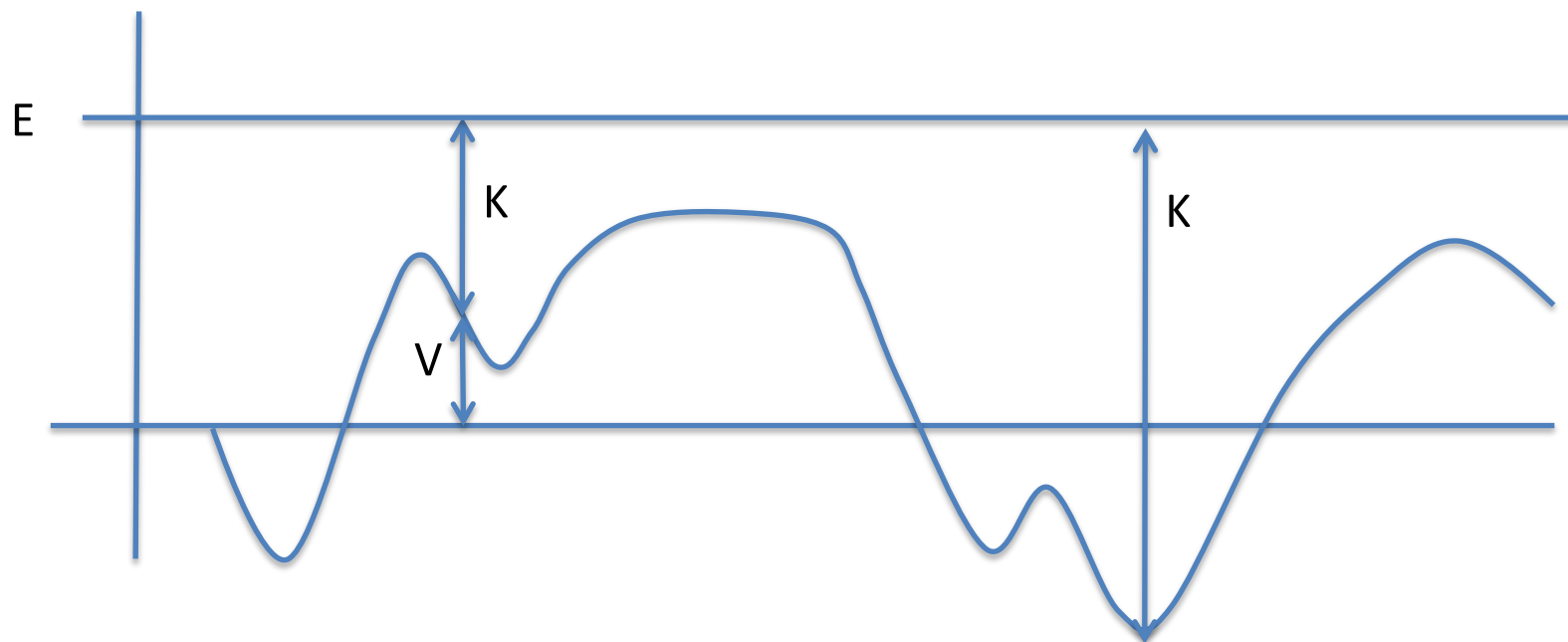
Conclusion: The approach helps to cross very high energy picks and reach conformations with minimum energy
Application: obtain structures with NMR data