Protein folding: some simple models

Henri Orland
SPhT, CEA-Saclay
France

work in collaboration with T. Garel

Outline

- What is a protein: chemistry, structure, interactions, energy scales, time scales, etc.
- The Hydrophobic effect: collapse vs. folding, entropy, ϑ-point
- Sequence diversity: heteropolymer models: the random bond model, the Hydrophobic-Hydrophilic model

- Dominant Folding Paths: Langevin dynamics, path integrals, dominant paths, Hamilton-Jacobi representation.

1. What is a Protein

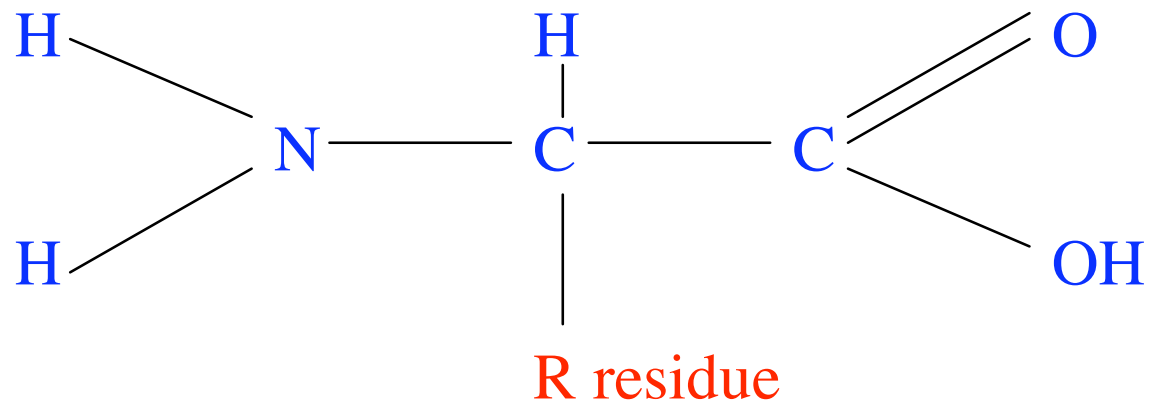
Biological Polymers (biopolymers): Proteins, Nucleic Acids (DNA and RNA), Polysaccharides

- catalytic activity: enzymes
- transport of ions: hemoglobin (O_2), ion channels
- motor protein
- shell of viruses (influenza, HIV, etc...)
- prions
- food, etc…

Proteins have an active site: biological activity

Polymers built with amino-acids

- 20 types of amino acids
- all left-handed
- **Ala, Ile, Leu, Met, Phe, Pro, Trp, Val, Asn, Cys, Gln, Gly, Ser, Thr, Tyr, Arg, His, Lys, Asp, Glu**
- 10 ≤ Number of Monomers ≤ 500



Among the 20 amino-acids:

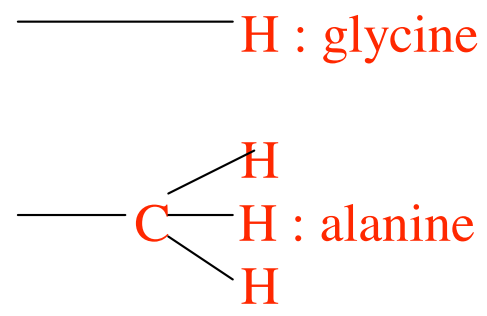
- 12 hydrophilic (polar)
- 8 hydrophobic (non polar)

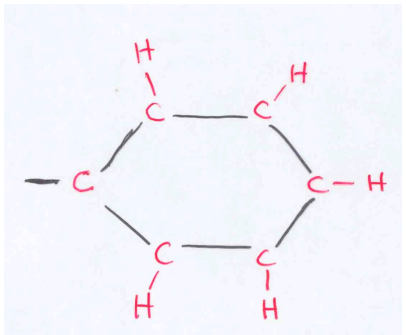
8 uncharged
4 charged

In a typical protein:

½ polar – ½ hydrophobic

Examples of residues:





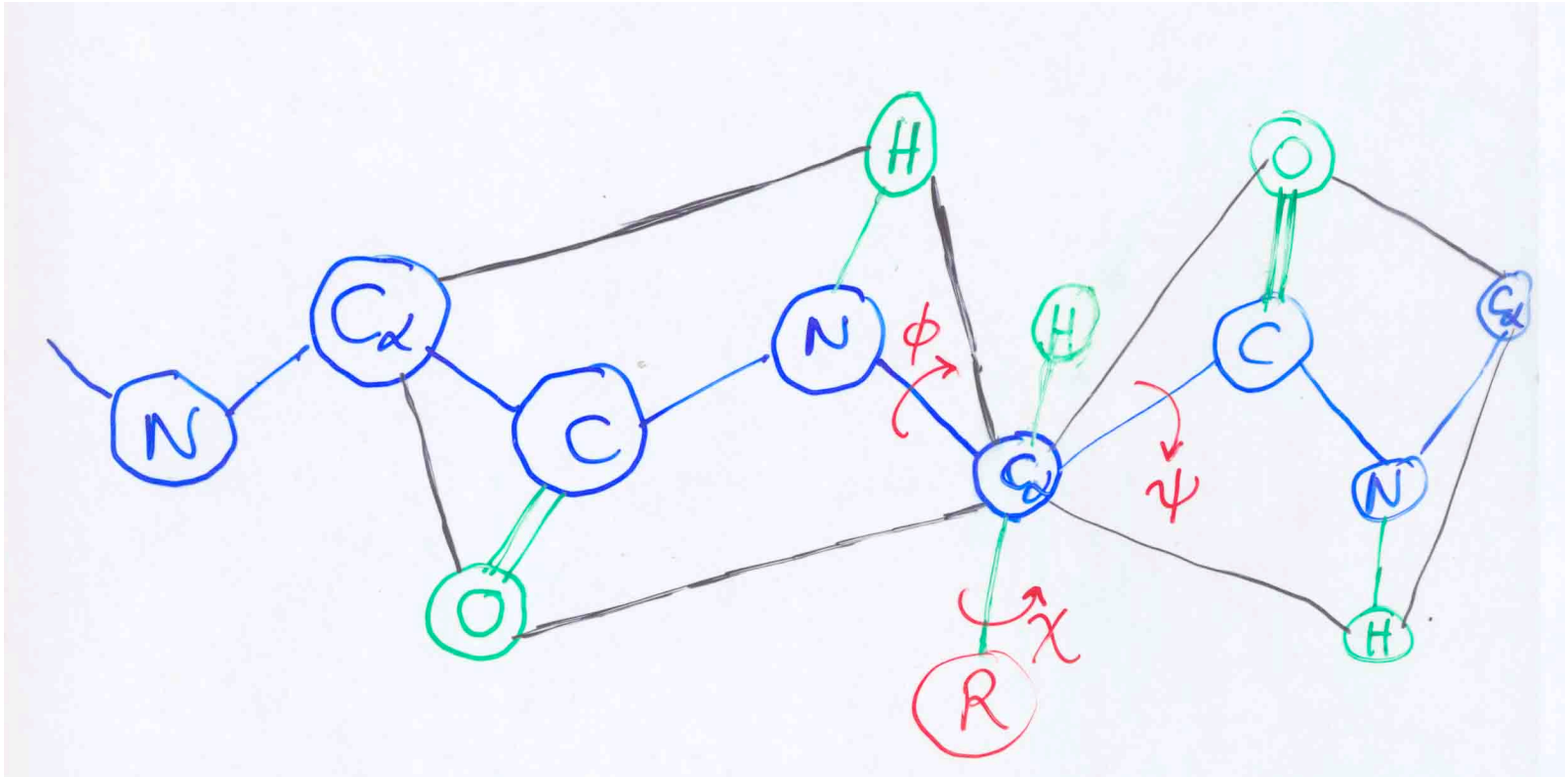
Polymerisation (polycondensation)

NH2---CHR1---COOH + NH2---CHR2---COOH + ...

NH2---CH—(CO—NH)-- CH—(CO—NH)-- CH—(CO—NH)---
| \----------/ | |
R1 peptide R2 R3
bond

+ H2O

⇒ weakly branched polymer



- Hard degrees of freedom:
  - ❖ covalent bonds
  - ❖ valence angles
  - ❖ peptide bonds
  - ❖ improper dihedrals

- Soft degrees of freedom
  - ♦ torsion angles : φ, ψ, χ very small energies

Proteins exist under two states:

♦ Denatured = Unfolded → Random Coil (swollen)
→ Molten Globule (compact)

No biological activity

♦ Native = Folded = Unique compact structure

Biologically active

Number of compact structures of a polymer :

$$\sim \mu^N$$

Puzzle: below folding transition temperature, the protein seems to exist under a unique conformation (zero conformational entropy).

Folding transition: depends on temperature, pH, denaturant agent, salt, etc…

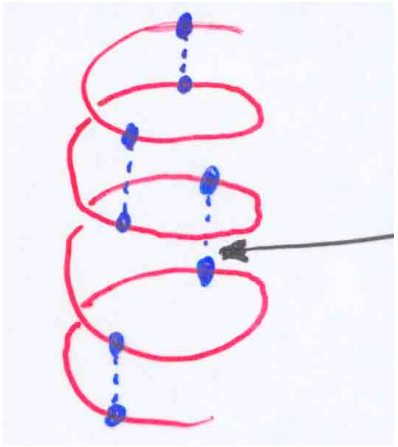
Time scales: Microscopic time : 10^{-15} s
Folding time: 10^{-2} to 1 s

Questions:

- Nature of the transition
  - crystallization (liquid-solid)
  - glass
  - purely dynamical
- How can one understand the uniqueness of the native state?
- Why is there so much secondary structure?
- What is the dynamics like? Exponential, stretched exponential, power law?

In all proteins, there is local order in the compact native state. These are the Secondary Structures.

- One dimensional: α-helix (mainly R)

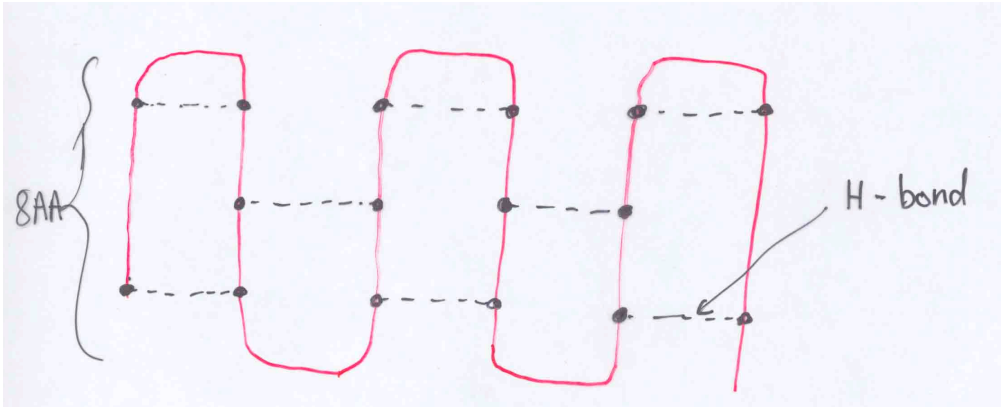


Hydrogen bonds

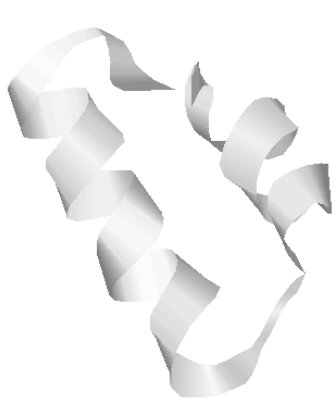
3.6 AA / turn

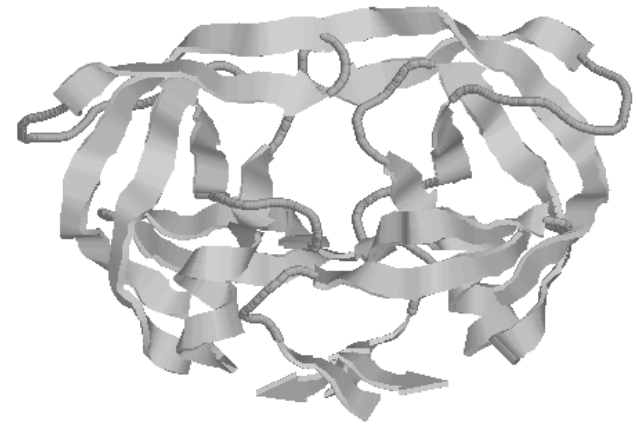
5-7 turns / helix

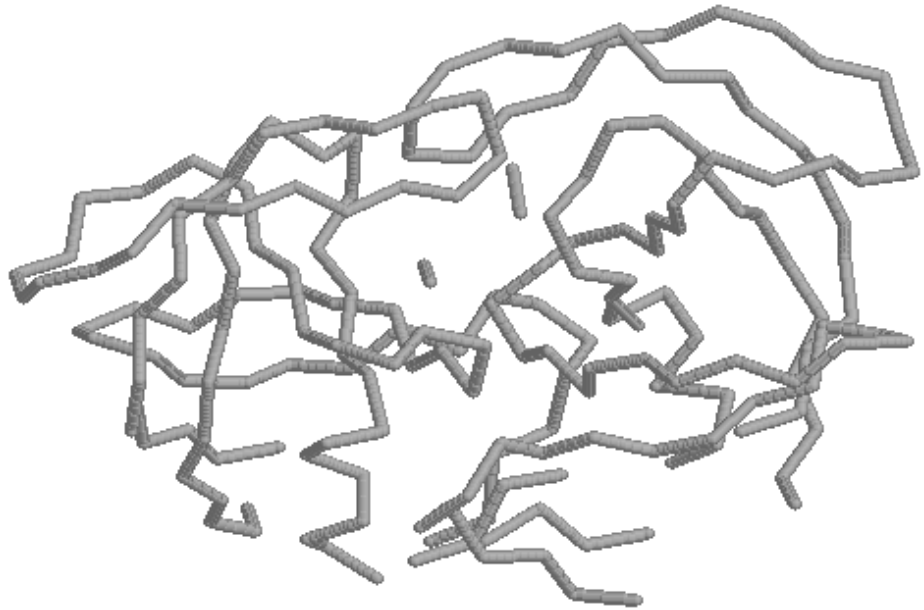
- Two dimensional: β-sheet



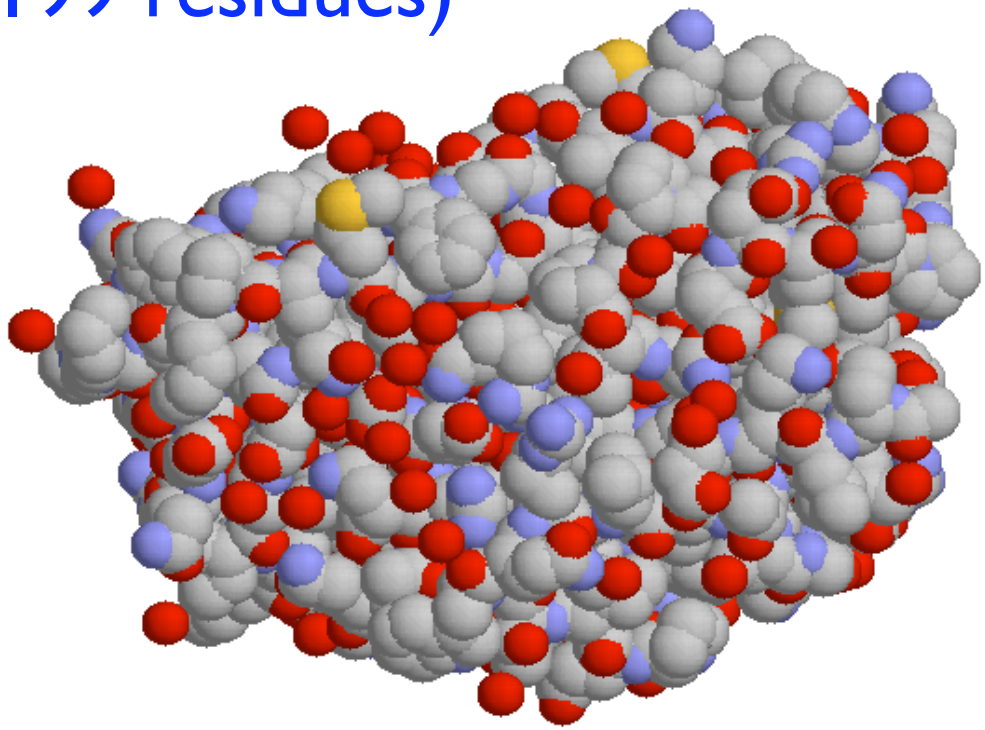
Tertiary structure: 3d structure of the folded protein→compact packing of secondary structures.







HIV protease (199 residues)



How can one see the folding transition?

1. Measure the Gyration radius: by X-rays or neutron scattering
2. Biological Activity
3. NMR
4. C.D (circular dichroism)

Renaturation time: 10^{-3} – 1 s

How can one see the 3d structure?

- X-ray Crystallography : but need to make a crystal first!
- NMR-NOE : measure nearby H-H pairs.

The Chemist's Approach

1. Look for effective atom-atom interactions → semi-empirical Hamiltonian
2. Molecular dynamics or Monte Carlo.

What interactions are present?

bonded
- -covalent bond
- -sulfur bridges (cysteins)

non bonded
- -Coulomb (with partial charges)
- -Van der Waals (steric repulsion)
- -Hydrogen bonds : intra-molecular or with the solvent.

The solvent is polar (Water) and induces hydrophobic interactions

which might be responsible for the collapse transition.

Energy Scales

1 eV = 23 kCal/mole = 10000° K

300° K = 0.6 kCal /mole

- Covalent bond: 50-150 kCal /mole
- Sulfur Bridge: 51 kCal/mole

- Hydrogen bonds: 5-8 kCal/mole (non polar solvent)
  1-2 kCal/mole (polar solvent)

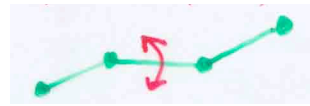
- Van der Waals: 1 kCal/mole
- Coulomb: 1-2 kCal/mole

Denaturation temperature ≈ 1 kCal/mole

Chemical sequence is frozen and only non-covalent interactions drive the folding.

Parametrization (CHARMM, AMBER, OPLS, …)






$$E = \sum_{bonds} k_b (b - b_0)^2 + \sum_{valence\ angles} k_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} k_\varphi (1 + \cos(n\varphi - \delta)) + \sum_{impropers} k_\nu (\nu - \nu_0)^2$$

$$+ \sum_{i<j} 4\varepsilon_{ij} \left((\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6} \right) + \sum_{i<j} \frac{332}{\varepsilon} \frac{q_i q_j}{r_{ij}}$$

Use Newton or Langevin dynamics

$$m_i \ddot{r}_i + \gamma_i \dot{r}_i + \frac{\partial E}{\partial r_i} = \eta_i(t)$$

where $\eta_i(t)$ is a Gaussian noise satisfying the fluctuation-dissipation theorem:

$$< \eta_i(t) \eta_j(t) >= 2\gamma_i k_B T \delta_{ij} \delta(t - t')$$

Then, it is well known that

$$P(\{r_i\},t) \xrightarrow[t\to\infty]{} \exp\left(-\frac{E(\{r_i\})}{k_B T}\right)$$

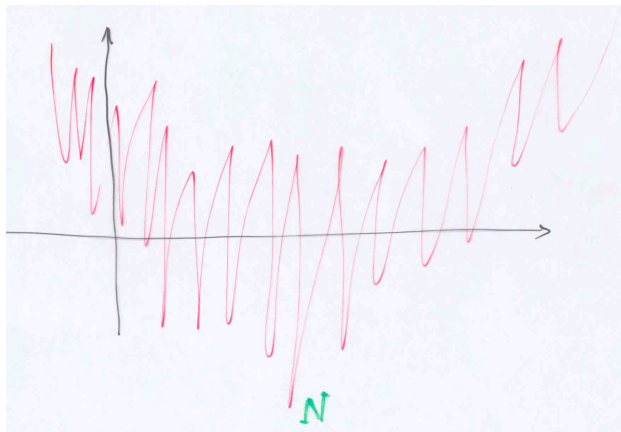
To discretize, one must use $\delta t \sim 10^{-15} - 10^{-13}$ s

Number of degrees of freedom: $N \geq 1000$

Longest available runs (with water) $t \sim 10^{-8}$ s

We see that $t \ll$ folding time.

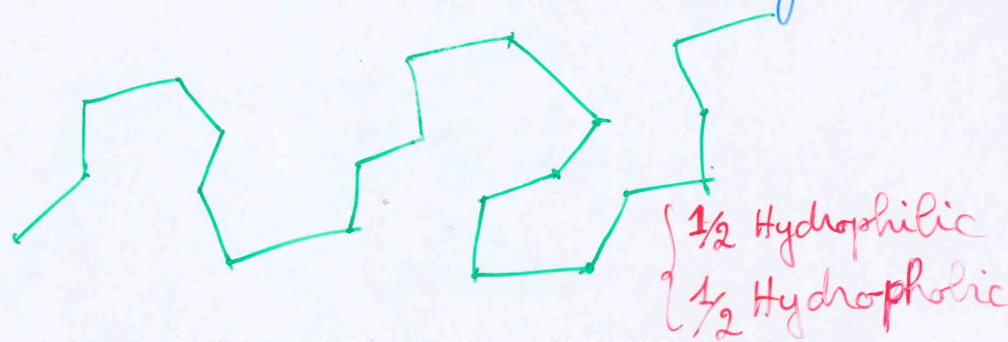
Reason: system is trapped in an exponential number of metastable traps.



The hydrophobic effect: Collapse transition

Simplified model:

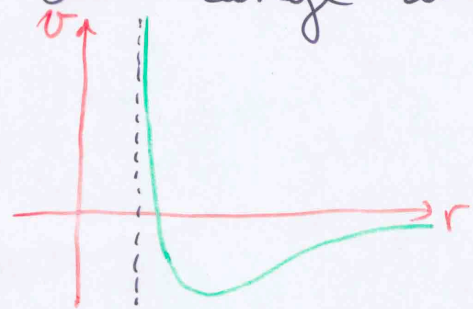
- one amino-acid = 1 monomer
  = 1 segment.



⟹ identical amino-acids, on the average hydrophobic.

Water induces attractive interaction between monomers: Polymer in a bad solvent (Flory)

Short range interactions



Lennard-Jones: $v(r) = \varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$

Partition function :

$$Z = \sum_{\{\text{chains } \vec{r}_i\}} e^{-\frac{\beta}{2} \sum_{i,j} v(\vec{r}_i - \vec{r}_j)}$$

Results : Θ - point (competition entropy $\leftrightarrow$ attraction)

Phase transition at $T = \Theta$,

between :

- swollen phase : $T > \Theta$

$$R_G \sim a N^{\nu} \quad , \quad \nu = \frac{3}{5}$$

- collapsed phase : $T < \Theta$

$$R_G \sim a N^{\nu} \quad , \quad \nu = \frac{1}{3}$$

At the Θ point, $R_G \sim a\sqrt{N}$

{ 2nd order transition
{ Tricritical point ($d_c = 3$)

- Can the collapsed state represent the folded state of the protein?
- No because collapsed state has finite conformational entropy (Hamiltonian walks)
- No because no secondary structure
- OK to describe the first stages of the folding transition ($t < 1\ \mu s$)

Sequence diversity: Heteropolymer models

- Very coarse model for the polymer

$$Z(\{v_{ij}\}) = \int \prod_i d\vec{r}_i \prod_i g(\vec{r}_i, \vec{r}_{i+1}) \ \exp(-\beta \mathcal{H}(\{v_{ij}\}))$$

- where $g(\vec{r}_i, \vec{r}_{i+1}) = \delta(|\vec{r}_i - \vec{r}_{i+1}| - a)$ or

- $$g(\vec{r}_i, \vec{r}_{i+1}) \rightarrow \exp(-\frac{d}{2a^2} \left(\frac{d\vec{r}(s)}{ds}\right)^2)$$

- Reduced Hamiltonian

$$\beta\mathcal{H}(\{v_{ij}\}) = \frac{1}{2}\sum_{i\neq j} v_{ij}(\vec{r}_i, \vec{r}_j) + \frac{1}{6}\sum_{i\neq j\neq k} w_0(\vec{r}_i, \vec{r}_j, \vec{r}_k) + ...$$

- The free energy is given by

$$F(\{v_{ij}\}) = -T\log Z(\{v_{ij}\})$$

- Assume 3-body term is not sequence dependent: excluded volume term

$$w_0(\vec{r}_i, \vec{r}_j, \vec{r}_k) = w\delta(\vec{r}_i - \vec{r}_j)\delta(\vec{r}_j - \vec{r}_k)$$

- Assume short-range sequence dependent two-body interaction:

$$v_{ij}(\vec{r}_i, \vec{r}_j) = v_{ij}\delta(\vec{r}_i - \vec{r}_j)$$

- To get an idea of the behavior of the system, assume quenched disorder:

$$F = \overline{F} = \int \prod P(\{v_{ij}\})F(\{v_{ij}\})d(\{v_{ij}\})$$

- "Average" heteropolymer

The random-bond heteropolymer

- It is defined by

$$v_{ij} = v_0 + w_{ij}$$

with w_{ij} random independent variables with distribution

$$h(w_{ij}) = \frac{1}{\sqrt{2\pi w^2}} \exp\left(-\frac{w_{ij}^2}{2w^2}\right)$$

- Average over Replicas

$$\overline{Z^n} = \int \prod_a \mathcal{D}\vec{r}_i^a \prod_{i,a} g(\vec{r}_i^a, \vec{r}_{i+1}^a) \exp\left(-\tilde{v}_0 \sum_{i<j}\sum_a \delta(\vec{r}_i^a - \vec{r}_j^a) + \frac{\beta^2 w^2}{2} \sum_{a\neq b}\sum_{i<j} \delta(\vec{r}_i^a - \vec{r}_j^a)\delta(\vec{r}_i^b - \vec{r}_j^b)\right)$$
$$\times \exp\left(-\frac{w_0}{6} \sum_{i\neq j\neq k}\sum_a \delta(\vec{r}_i^a - \vec{r}_j^a)\ \delta(\vec{r}_j^a - \vec{r}_k^a)\right)$$

with $\tilde{v}_0 = v_0 - \beta^2 \frac{w^2}{2}$.

- Introduce density order parameters and overlap order parameters (and conjugate)

$$\overline{Z^n} = \int \mathcal{D}q_{ab}(\vec{r},\vec{r}')\mathcal{D}\hat{q}_{ab}(\vec{r},\vec{r}')\mathcal{D}\rho_a(\vec{r})\mathcal{D}\phi_a(\vec{r}) \exp\left(G(q_{ab},\hat{q}_{ab},\rho_a,\phi_a) + \log\zeta(\hat{q}_{ab},\phi_a)\right)$$

$$\rho_\alpha(\vec{r}) = \sum_{i=1}^N \delta(\vec{r}_i^{(\alpha)} - \vec{r})$$

$$q_{\alpha,\beta}(\vec{r},\vec{r}') = \sum_{i=1}^N \delta(\vec{r}_i^{(\alpha)} - \vec{r})\delta(\vec{r}_i^{(\beta)} - \vec{r}')$$

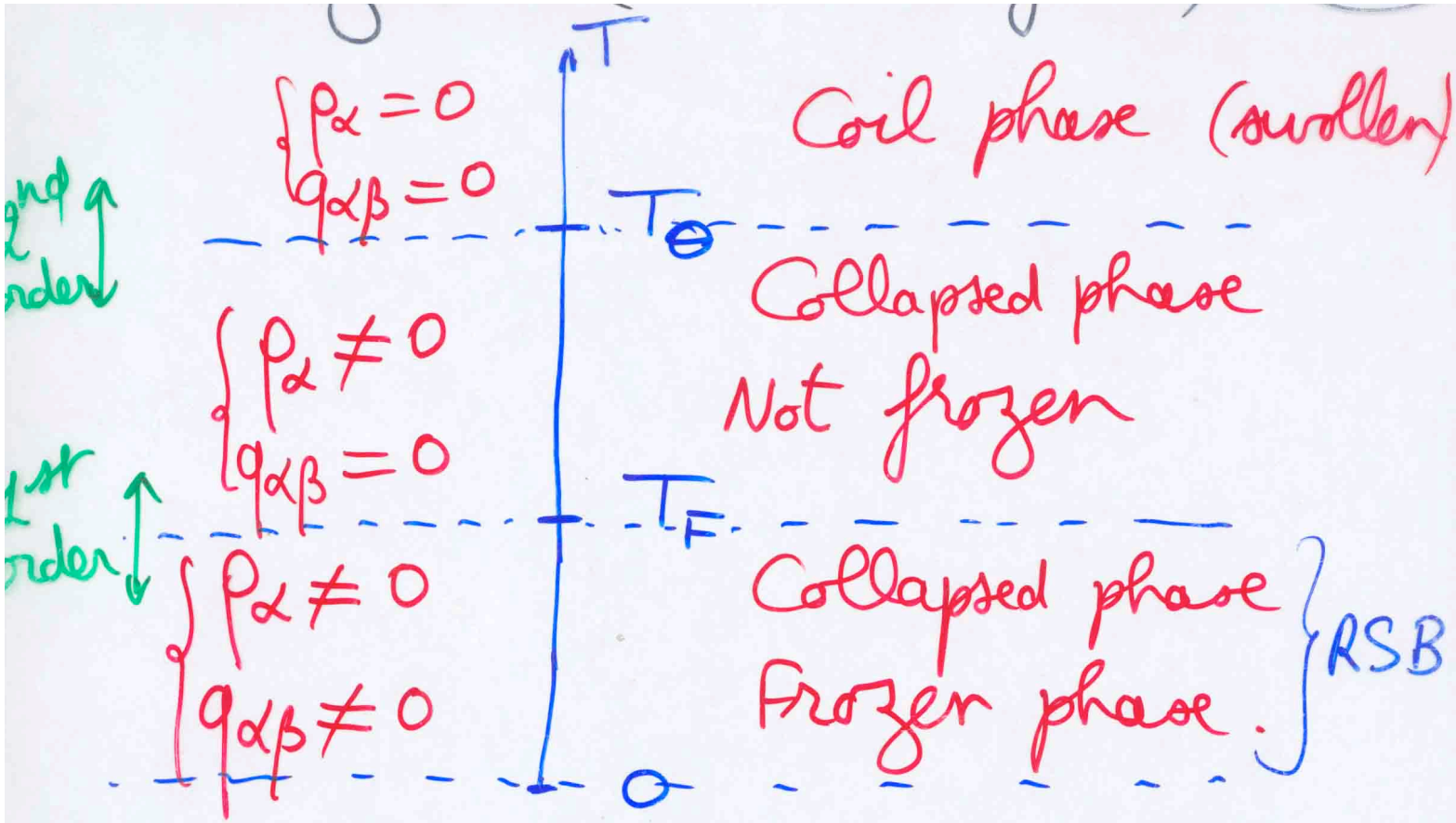
$$G(q_{ab},\hat{q}_{ab},\rho_a,\phi_a) = \int d^d r \sum_a \left(i\rho_a(\vec{r})\phi_a(\vec{r}) - (\tilde{v}_0)\ \frac{\rho_a^2(\vec{r})}{2} - \frac{w_0}{6}\rho_a^3(\vec{r}) \right)$$
$$+ \int d^d r \int d^d r' \sum_{a<b} \left(i q_{ab}(\vec{r},\vec{r}')\hat{q}_{ab}(\vec{r},\vec{r}') + \frac{\beta^2 w^2}{2}\ q_{ab}^2(\vec{r},\vec{r}') \right)$$

$$\zeta(\hat{q}_{ab},\phi_a) = \int \mathcal{D}\vec{r}_a(s) \exp\left(-\frac{d}{2a^2} \int_0^N ds\ \dot{\vec{r}}_a^{\,2} \right)$$
$$\times\ \exp\left(-i\int_0^N ds\ \sum_a \phi_a(\vec{r}_a(s)) - i\int_0^N ds\ \sum_{a<b} \hat{q}_{ab}(\vec{r}_a(s),\vec{r}_b(s)) \right)$$

- at high temperature, $\rho_\alpha = 0$ and $q_{\alpha,\beta} = 0$ It is the Denatured phase (SAW)

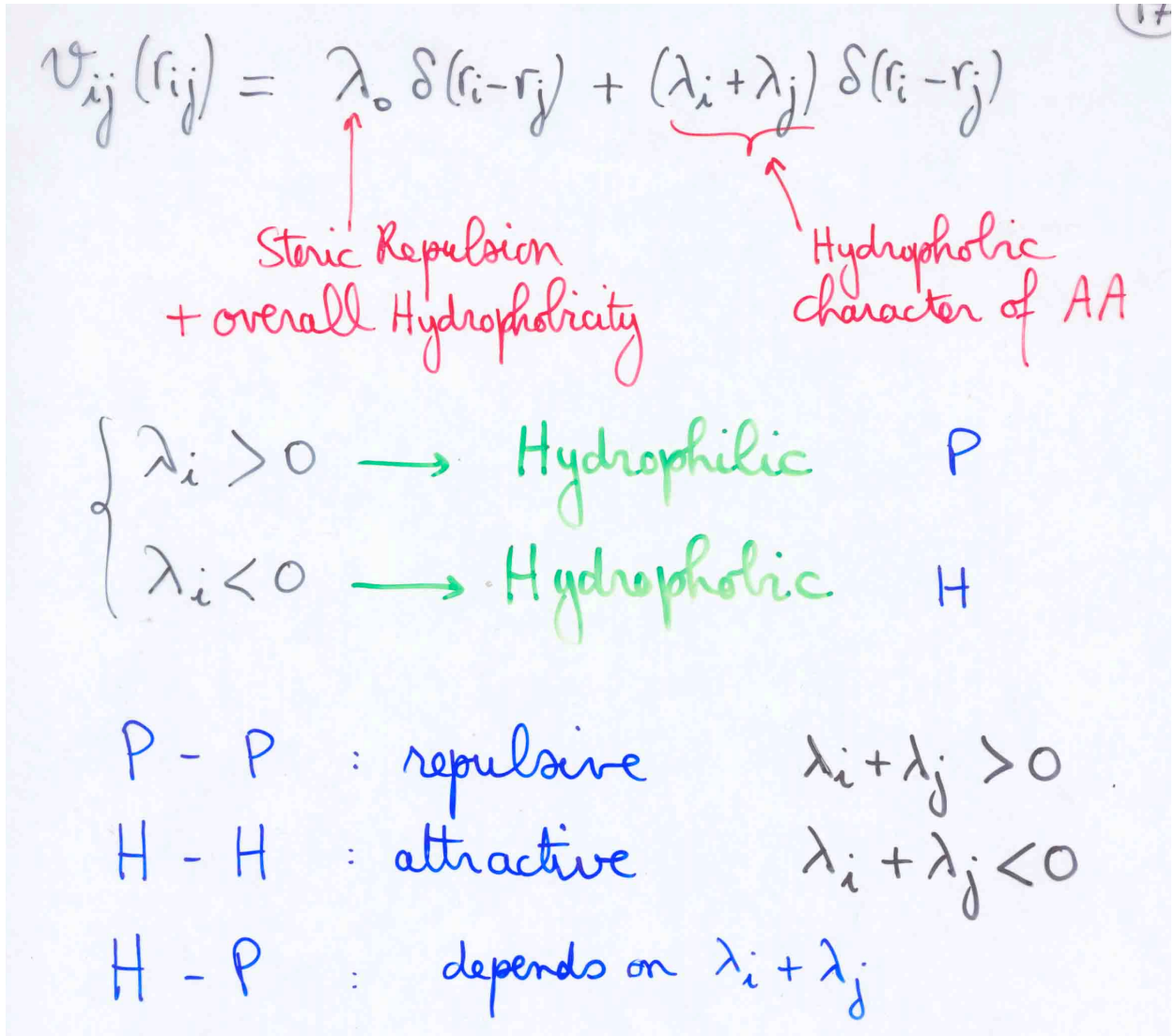
- at lower temperature, T_θ , collapsed phase with $\rho_\alpha \neq 0$ and $q_{\alpha,\beta} = 0$ Molten Globule

- at lower temperature, T_F , freezing transition $\rho_\alpha \neq 0$ and $q_{\alpha,\beta} \neq 0$ Native?



Frozen phase is similar to low T phase of a spin glass (Parisi-like RSB) $\Longrightarrow$ existence of few dominant states. Native state?

The Hydrophobic-Hydrophilic chain



Take λ_i random independent

$$P(\lambda_i) = \frac{1}{\sqrt{2\pi\lambda^2}} e^{-\frac{\lambda_i^2}{2\lambda^2}}$$

$$Z = \sum_{\{\text{chains}\}} e^{-\frac{\beta}{2}\sum_{i,j}\left[\lambda_0 \delta(\vec{r}_i - \vec{r}_j) + (\lambda_i + \lambda_j)\,\delta(\vec{r}_i - \vec{r}_j)\right]}$$
$$e^{-\frac{\beta}{6} w_3 \sum_{i,j,k} \delta(\vec{r}_i - \vec{r}_j)\,\delta(\vec{r}_j - \vec{r}_k)}$$
$$e^{-\frac{\beta}{24} w_4 \sum_{i,j,k,\ell} \delta(\vec{r}_i - \vec{r}_j)\,\delta(\vec{r}_j - \vec{r}_k)\,\delta(\vec{r}_k - \vec{r}_\ell)}$$

Do Quenched Average over the λ_i

$$\left\langle e^{-\beta \sum_{i=1}^{N} \lambda_i \sum_{\alpha=1}^{n} \sum_{j=1}^{N} \delta\left(\vec{r}_i^{(\alpha)} - \vec{r}_j^{(\alpha)}\right)} \right\rangle$$

$$= e^{\frac{\beta^2 \lambda^2}{2} \sum_{\alpha,\beta=1}^{n} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \delta\left(\vec{r}_i^{(\alpha)} - \vec{r}_j^{(\alpha)}\right) \delta\left(\vec{r}_i^{(\beta)} - \vec{r}_k^{(\beta)}\right)}$$

↓

attractive 3-body term

↳ need a repulsive 4th virial

Effective 3rd virial:

$$w_3' = w_3 - 3\beta\lambda^2 = w_3 - \frac{3\lambda^2}{T}$$

By introducing order parameters:

$$\begin{cases} \rho_\alpha(\vec{r}) = \sum_{i=1}^{N} \delta(\vec{r} - \vec{r}_i^{(\alpha)}) \\ q_{\alpha\beta}(\vec{r},\vec{r}') = \sum_{i=1}^{N} \delta(\vec{r} - \vec{r}_i^{(\alpha)})\, \delta(\vec{r}' - \vec{r}_i^{(\beta)}) \end{cases} , \quad \alpha = 1,\dots,m$$

(176)

$$\overline{Z^n} = \int \mathcal{D}\rho_\alpha(r)\, \mathcal{D}\hat{\rho}_\alpha(r)\, \mathcal{D}q_{\alpha\beta}(r,r')\, \mathcal{D}\hat{q}_{\alpha\beta}(r,r')$$

$$e^{i\sum_\alpha \int dr\, \hat{\rho}_\alpha(r)\rho_\alpha(r) + i\sum_{\alpha\beta}\int dr\,dr'\,\hat{q}_{\alpha\beta}(r,r')\,q_{\alpha\beta}(r,r')}$$

$$e^{-\frac{\lambda_0}{2}\sum_\alpha \int dr\,\rho_\alpha^2(r) - \frac{w_3}{6}\sum_\alpha \int dr\,\rho_\alpha^3(r) - \frac{w_4}{24}\sum_\alpha \int dr\,\rho_\alpha^4(r)}$$

$$e^{\frac{\beta^2\lambda^2}{2}\sum_{\alpha,\beta}\int dr\,dr'\,\rho_\alpha(r)\,q_{\alpha\beta}(r,r')\,\rho_\beta(r') + \mathrm{Log}\,\mathcal{U}(\hat{\rho}_\alpha,\hat{q}_{\alpha\beta})}$$

with

$$\mathcal{U}(\hat{\rho}_\alpha,\hat{q}_{\alpha\beta}) = \int \mathcal{D}r_\alpha(s)\, e^{-\frac{3}{2a^2}\sum_\alpha \int_0^N ds \left(\frac{d\vec{r}_\alpha}{ds}\right)^2 - i\sum_\alpha \int_0^N ds\, \hat{\rho}_\alpha(\vec{r}_\alpha(s))}$$

$$\times e^{-i\sum_{\alpha,\beta}\int_0^N ds\, \hat{q}_{\alpha\beta}(\vec{r}_\alpha(s),\vec{r}_\beta(s))}$$

Because of the structure of overlap terms, no replica symmetry breaking

$\Rightarrow$ no glass transition

$$i\hat{q}_{\alpha\beta}(r,r') = -\frac{\beta^2\lambda^2}{2}\rho_\alpha(r)\rho_\beta(r')$$

Approximations

* Mean Field theory
* Ground State Dominance

Results

* for any λ_0 (overall hydrophobicity)

$\longrightarrow$ "collapse transition"

* no Replica Symmetry Breaking $\longrightarrow$

Not a freezing glass transition.

$\lambda_0 < 0$

Strongly Hydrophobic case

Second order transition, driven by the strong $\lambda_0 < 0$ term

Ordinary Θ point

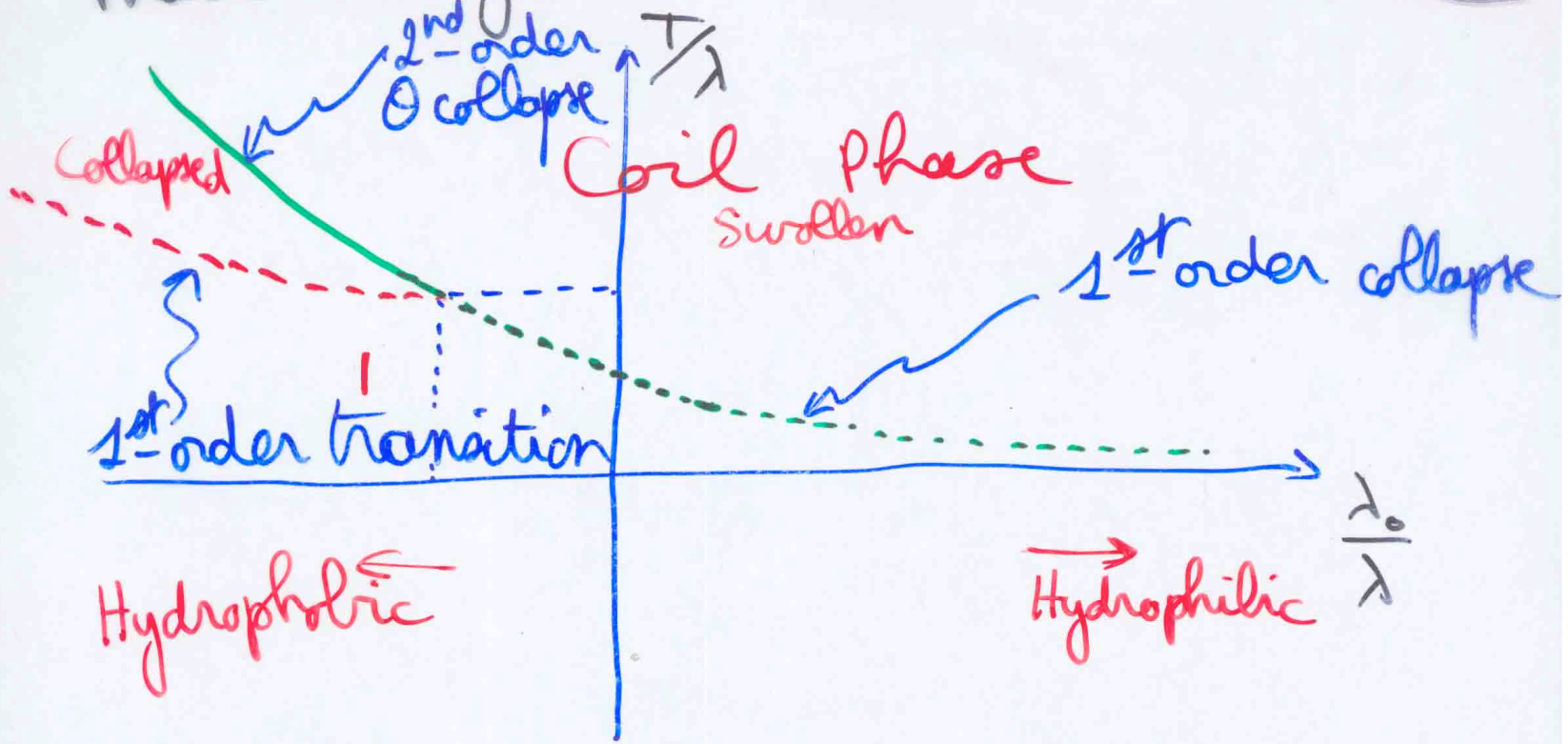
Weakly Hydrophobic or Hydrophilic case

First order "collapse" transition, induced by the disorder fluctuations of the 3-body term.

- Not an ordinary Θ-point
- Not a glass transition.

1st-order ⟹
- * metastability and retardation effects
- * latent heat ⟹ strong entropy reduction

Phase diagram

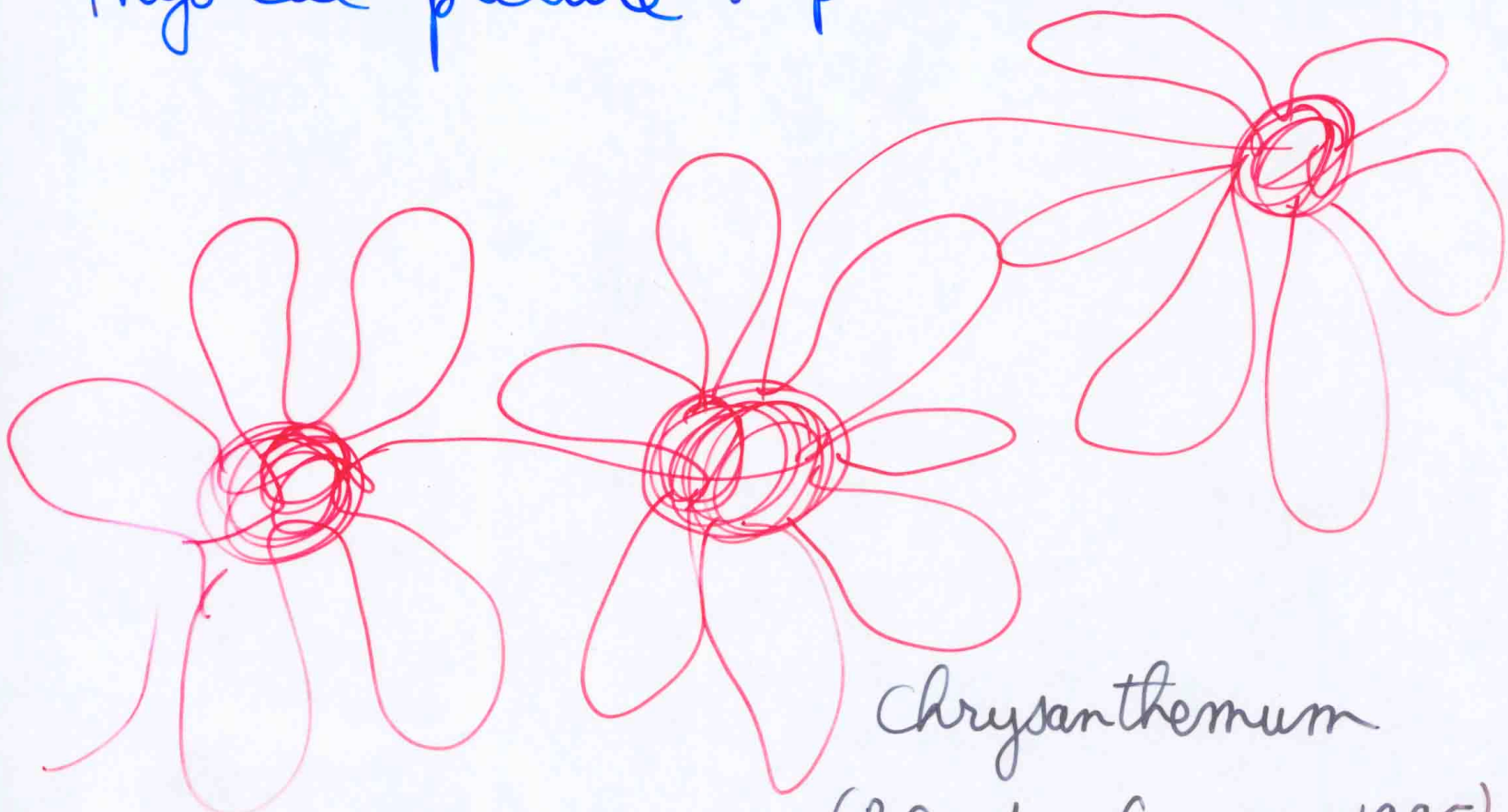


1st order transition ⇒ latent heat

⇒ reduction of entropy.

However, no unique ground state!

Physical picture : phase coexistence



Chrysanthemum

(P.G de Gennes, 1995)

Dominant Folding Pathways

- The problem: Assume a protein can go from state A to state B. Which pathway (or family of pathways) does the protein take?
- Examples:
  - from denatured to native in native conditions
  - Allosteric transition between A and B

with P. Faccioli, F. Pederiva and M. Sega

The case of one particle

- Take Langevin (Brownian) dynamics

$$\frac{\partial x}{\partial t} = -\frac{D}{k_B T}\frac{\partial U}{\partial x} + \eta(t)$$

- with Gaussian noise: $\langle \eta(t)\eta(t') \rangle = 2D\delta(t-t')$

- The Probability to find the particle at x at time t is given by a Fokker-Planck equation

$$\frac{\partial}{\partial t} P(x,t) = D\frac{\partial}{\partial x}\left(\frac{1}{k_B T}\frac{\partial U(x)}{\partial x} P(x,t)\right) + D\frac{\partial^2}{\partial x^2} P(x,t)$$

- Stationary distribution: the Boltzmann distribution

$$P(x) \sim \exp(-U(x)/k_B T)$$

- General form:

$$P(x_f, t_f | x_i, t_i) = e^{-\frac{U(x_f)-U(x_i)}{2k_B T}} \int_{x_i}^{x_f} \mathcal{D}x(\tau)\, e^{-S_{eff}[x]/2D}$$

- Boundary conditions:

$$x(t_i) = x_i \qquad\qquad x(t_f) = x_f$$

- The effective action is given by

$$S_{eff}[x] = \int_{t_i}^{t} d\tau \left(\frac{\dot{x}^2(\tau)}{2} + V_{eff}[x(\tau)] \right)$$

- and the effective potential is given by

$$V_{eff}(x) = \frac{D^2}{2} \left(\frac{1}{k_B T} \frac{\partial U(x)}{\partial x} \right)^2 - \frac{D^2}{k_B T} \frac{\partial^2 U(x)}{\partial x^2}$$

- Dominant trajectories: classical trajectories

$$\frac{d^2x}{dt^2} = -\frac{\partial(-V_{eff}[x])}{\partial x}$$

- with correct boundary conditions.

- Problem: one does not know the transition time.

- Solution: go from time-dependent Newtonian dynamics to energy-dependent Hamilton-Jacobi description.

- The method: minimize the Hamilton-Jacobi action

$$S_{HJ} = \int_{x_i}^{x_f} dl \sqrt{2\left(E_{eff} + V_{eff}[x(l)]\right)}$$

- over all paths joining x_i to x_f

dl is an infinitesimal displacement along the path

E_{eff} is a free parameter

- The total time of passage is determined by
-

$$t_f - t_i = \int_{x_i}^{x_f} dl \sqrt{\frac{1}{2\left(E_{eff} + V_{eff}[x(l)]\right)}}.$$

- E_{eff} is not the true energy of the system
- If the final state is an equilibrium state, then

$$E_{eff} = -V_{eff}(x_f)$$

- the Villin Headpiece Subdomain

- The HJ method is much more efficient than Newtonian mechanics because proteins spend most of their time trying to overcome free-energy barriers.

- No waiting-times in HJ: work with fixed interval length dl

- For a Protein, minimize

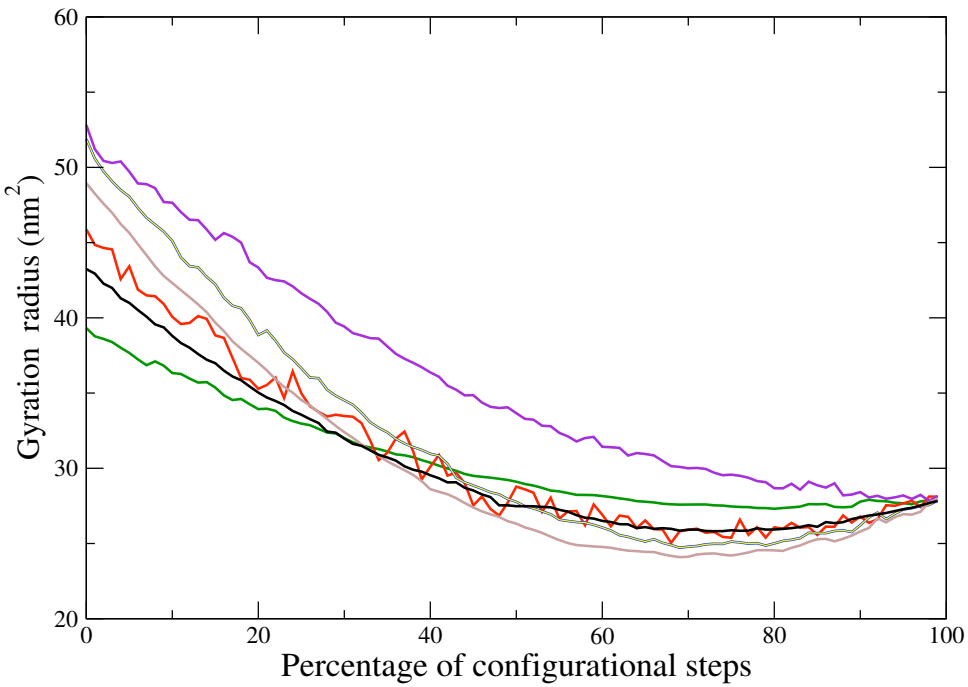
$$S_{HJ} = \sum_{n}^{N-1} \sqrt{2\left(E_{eff} + V_{eff}(n)\right)}\Delta l_{n,n+1} + \lambda P$$

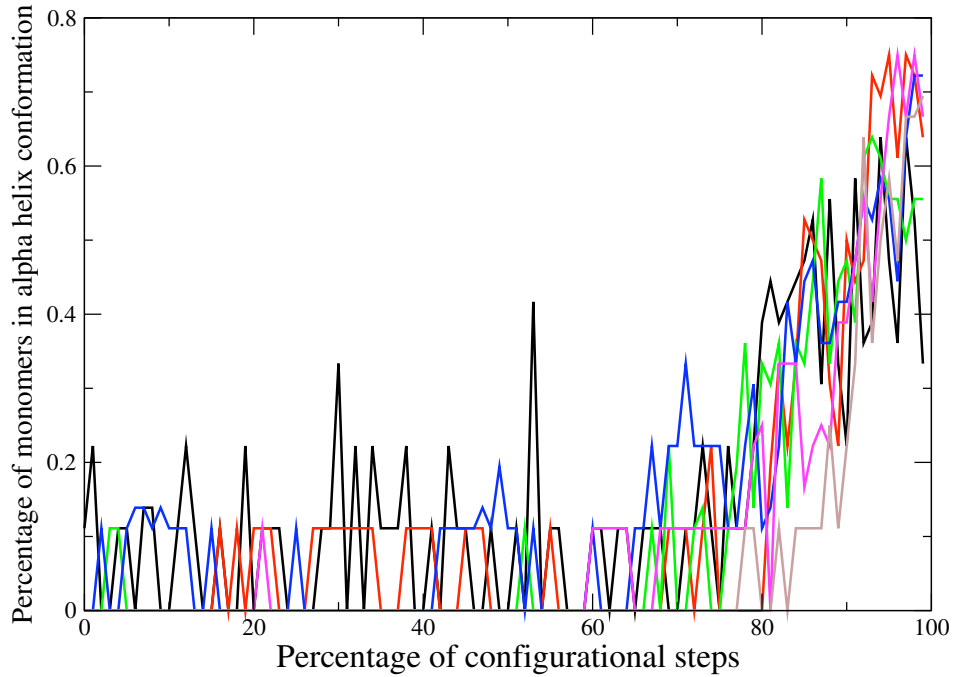
- where $P = \sum_{i}^{N-1}(\Delta l_{i,i+1} - \langle \Delta l \rangle)^2$ and λ is a Lagrange multiplier to fix the interval length

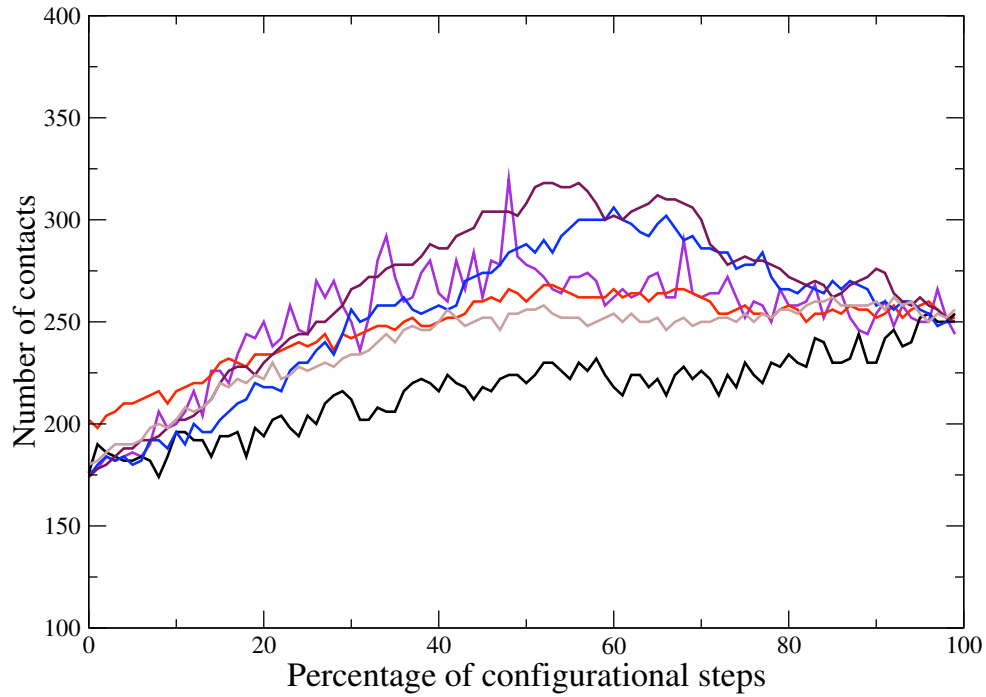
$$\begin{aligned} V_{eff}(n) &= \sum_{i} \left[\frac{D^2}{2(k_B T)^2} \left(\sum_{j} \nabla_j u(\mathbf{x}_i(n), \mathbf{x}_j(n)) \right)^2 \right. \\ &- \left. \frac{D^2}{k_B T} \sum_{j} \nabla_j^2 u(\mathbf{x}_i(n), \mathbf{x}_j(n)) \right] \\ (\Delta l)^2_{n,n+1} &= \sum_{i} (\mathbf{x}_i(n+1) - \mathbf{x}_i(n))^2, \end{aligned}$$

Results for the Villin Headpiece

Go Model




Percentage of monomers in alpha helix conformation
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100
Percentage of configurational steps


Number of contacts
400
350
300
250
200
150
100
0
20
40
60
80
100
Percentage of configurational steps