RNA Statistical Mechanics

Enzo Marinari

(SMC-INFM and Roma *La Sapienza*)

Among many other relevant works:

- A. Pagnani, F. Ricci-Tersenghi, EM, cond-mat/0111172;
- A. Pagnani, G. Parisi, F. Ricci-Tersenghi, cond-mat/9907125;
- R. Bundschuh and T. Hwa, cond-mat/0107210;
- F. Krzakala, M. Mézard e M. Müller, cond-mat/0108374.

Montpelier, March 2002

Statistical Mechanics

of

Byopolymers:

the Role of Disorder

Can Statistical Mechanics of Disordered Systems play a role in this game?

Large number of components.

Quenched Disorder.

But: not clear a priori.

I will try to discuss low T properties of models that can be relevant to the description of secondary structure of RNA.

RNA is a fundamental bio-polymer crucial for all living systems.

4 bases: A, C, G and U (like for DNA).

Double helical structures $\longrightarrow$ stacks of stable Watson-Crick pairs.

But: DNA is double stranded (perfect complementarity, wonderful way for storing information). Tertiary structure of double helical parts.

RNA: usually single stranded.

Two different phases:

1. formation of the secondary structure. coupling of basis.
2. formation of the tertiary structure.

Time scale separation.

Experimental side.

Precise, single molecule experiments.

Different experimental methods.

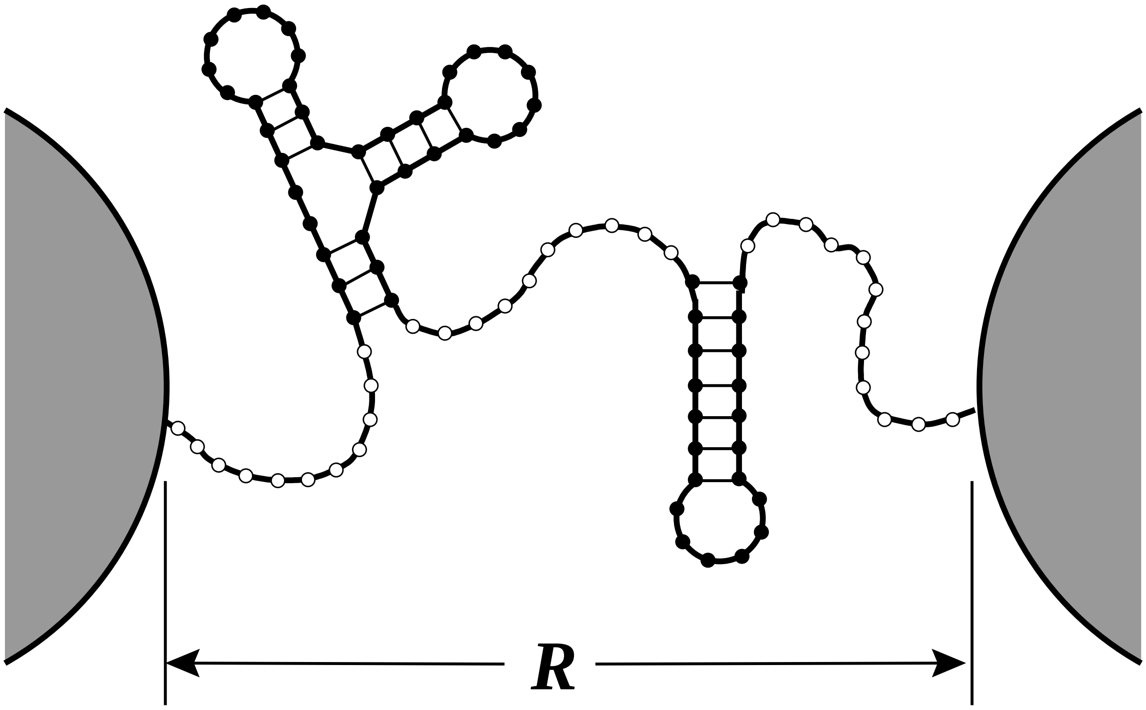
⟶ elastic properties of the polymer (persistence length, longitudinal elasticity);

⟶ structural unfolding transition.

via

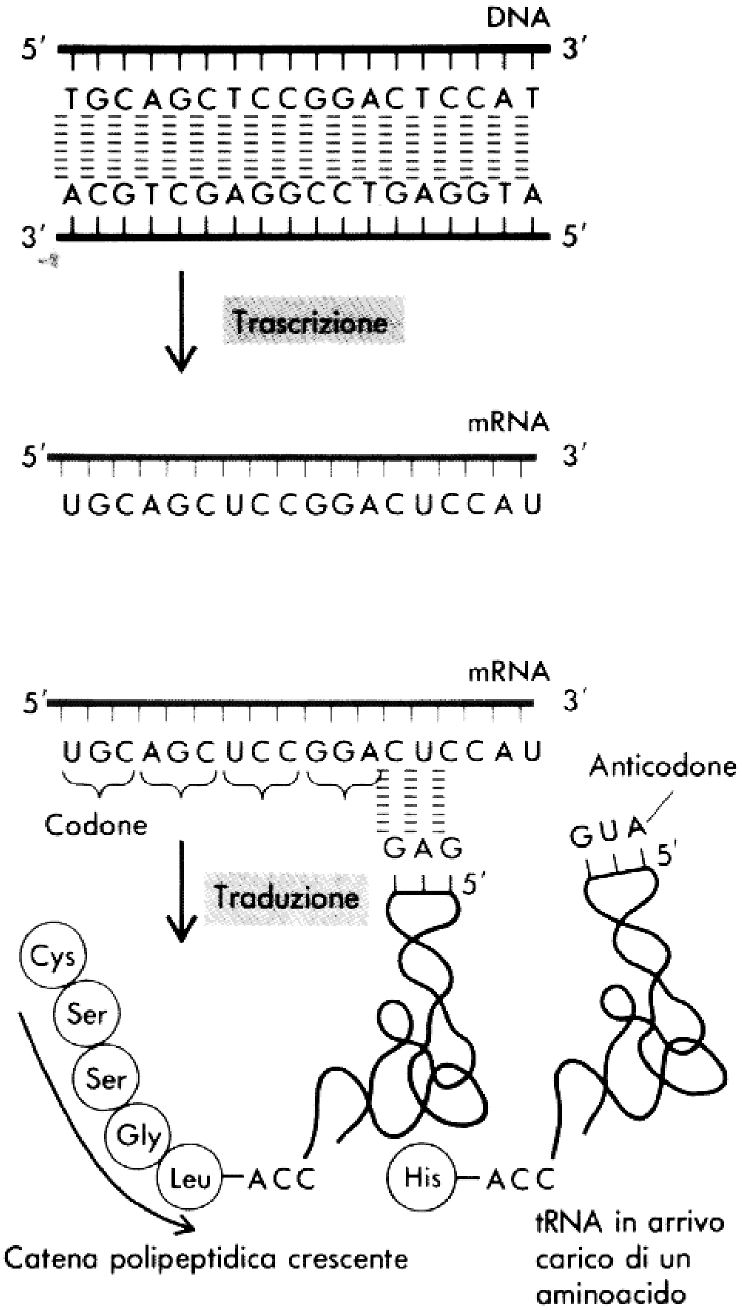
force-extension curves

"unzipping" of double stranded DNA.

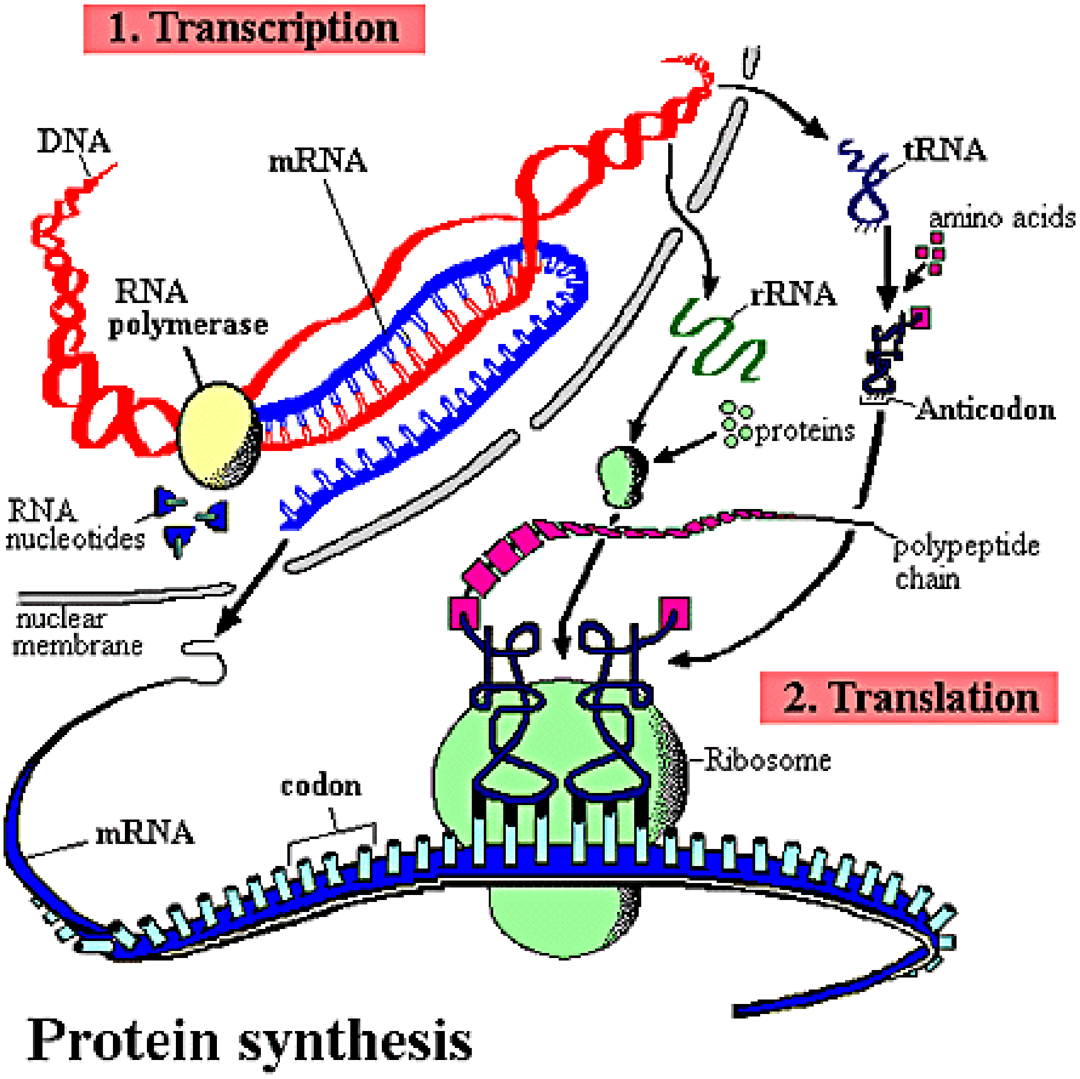


from Garland, Bundschuh and Hwa, cond-mat/0101250

The Transcription of RNA



The Ribosomal Cycle



Different types of RNA.

1. **Messenger RNA: mRNA** $\simeq 1000, 2000$ bases long. Its length is very variable, and can range up to 6800 bases. It is the carrier of the genetic information about primary structure of proteins from DNA, together with special features that allow it to attach to ribosomes and function in protein synthesis. Its size depends on the size of the protein for which it codes. It tends to be relatively short-lived, and its lifetime varies from molecular species to molecular species (depending on the biological role of the protein).

2. **Transfer RNA: tRNA** $\simeq 70, 100$ bases long. It is designed to carry activated amino acids to the site of protein synthesis, the ribosome. It is long-lived (stable).

3. **Ribosomal RNA: rRNA** $\simeq 120, 4700$ bases long. It forms the ribosome, the site of protein synthesis. Various species range in size from 4700 bases to about 120 bases. Eukaryotic and prokaryotic rRNA's are distinctly different. rRNA is also long-lived (stable).

General points that I will try to stress and follow (ideas and things that can be useful in the study of "disordered" and "complex" systems):

1. ϵ-coupling method (here and, in general, as applied to the study of disordered and glassy systems).
2. degeneration (we study different models). (see for example scaling properties of 2D Spin Glasses)
3. glassy phase

I will describe some recent work (A. Pagnani, F. Ricci-Tersenghi, EM) that I will also use as a guide.

We expect a glassy phase to be present (Bundschuh-Hwa) at low T.

Here:

scaling regime

of

low energy excitations

in

different models

of

random RNA.

Secondary structures.

ϵ-coupling method (see also Krzakala, Mézard and Müller).

We want to compute the energy cost of typical excitations above the ground state

(they can be "large", i.e. involving a finite fraction of the system basis).

$$\Delta E(N) \propto N^{\theta}$$

N: number of basis in the chain.
θ: relevant scaling exponent.

(other methods can be used: they would probably give the same results, see later during this talk)

$\longrightarrow$ interesting question about the universality of the concept of typical excitations (Picco, Ritort and Sales);
$\longrightarrow$ our bulk perturbation: surely very nice.

Our perturbation: repulsion from a given ground state structure (the first GS we find).

Secondary structure of RNA:

$\rightarrow$ set of base pairs that occur in its $3D$ structure.

Sequence of basis:

$\rightarrow$ $\mathcal{R} \equiv \{r_i\} = (r_1, r_2, \cdots, r_N)$

$\rightarrow$ r_i: i^{th} basis on the chain

$\rightarrow$ $r_i \in \{A, C, G, U\}$ four letter alphabet.

Secondary structure on $\mathcal{R}$: set $\mathcal{S}$ of (i, j) pairs.

$$1 \leq i < j \leq N$$

Every basis can appear in $\mathcal{S}$ 0 or 1 times.

Additional rules for $\mathcal{S}$.

First is crucial: **Planarity condition**

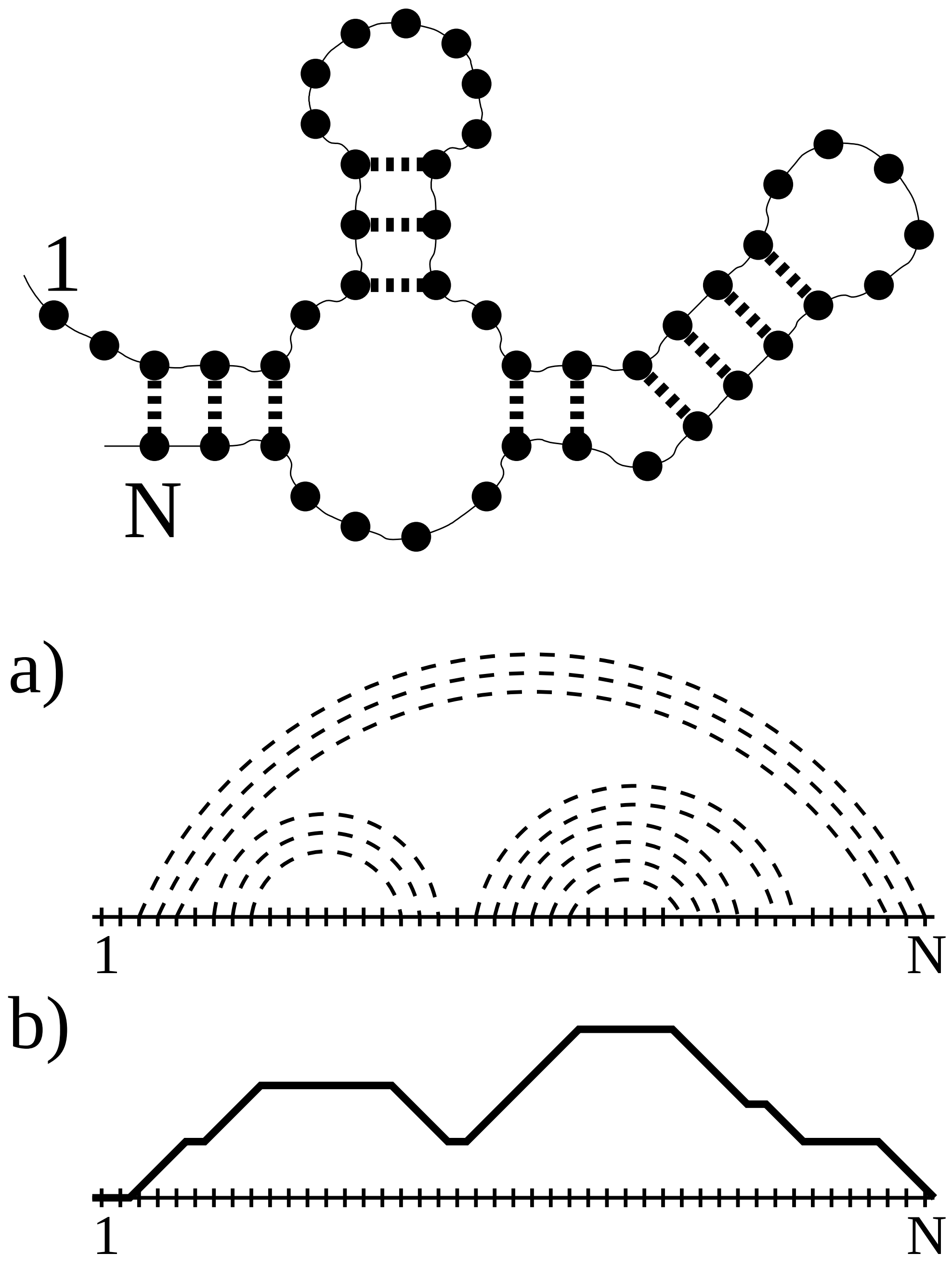
(i, j), (i', j') can belong to $\mathcal{S}$ (let without loss of generality be $i < i'$)

iff

$i < j < i' < j'$ or $i < i' < j' < j$

Excludes pseudo-knots. Very rare in real DNA.

The planarity condition is crucial for analytic and numerical approaches to the problem.



are equivalent representation of the RNA string.

More rules can be implemented in order to get closer to the experimental situation.

In our work for example we impose

$$j - i \geq 4$$

that allows a better flexibility of the chain in its $3D$ arrangement.

The Hamiltonian of our RNA model is:

$$\mathcal{H} = \sum_{(i,j)} e_{i,j} = \sum_{i,j} e_{i,j}\ l_{i,j}$$

$e_{i,j}$: "quenched disorder", fixed, i.e. pairing energy among basis i and j (Watson-Crick).

$e_{i,j} = e(r_i, r_j)$, E(C-G), E(A-U) at room temperature are of the order of one kcal/mole.

$l_{i,j}$: dynamical variables. $l_{i,j} = 1$ if $(i, i) \in \mathcal{S}$, 0 otherwise.

In better detail we use:

4 symbols A, C, G and U.

Only Watson-Crick couples:

C-G "strong" -2

A-U "weak" -1

All the other couplings are positive so they never appear.

This model has an exponentially

large degeneracy of the GS

$\Longrightarrow$ pathology (this is because we have allowed only two possible values for the energy).

Accidental degeneracy, very probably not relevant in an experimental framework. The real energy is a very complex object, that includes a large number of interactions.

Because of that we are also interested in studying models without degeneration.

Two different ways for removing the degeneration. In both cases

$$e'_{ij} = e_{ij} + \eta_{ij}$$

η_{ij}: small perturbing term. Quenched random. Depending on r_{ij}.

QD-model (quasi-degenerate)

$$\langle \eta \rangle = 0, \quad \langle \eta^2 \rangle = \frac{{\eta_0}^2}{L}$$

where η_0 is a small constant, $O(0.1)$.

When $L \to \infty$ energies are modified of an infinitesimal amount.

$\langle \eta^2 \rangle$ is such that the energies of the GS (that are degenerate for $\eta_0 = 0$) are split over a range $O(\eta_0)$.

Unphysical degeneracy is lifted.

The structure of the original spectrum is preserved.

Sequence still plays a role.

ND-model (non-degenerate)

$$\langle \eta \rangle = 0, \quad \langle \eta^2 \rangle = {\eta_0}^2 (\eta_0 = 0.1)$$

variance is finite when $L \to \infty$. It induces an $O(\sqrt{L})$ splitting of the GS that were degenerate in the **D**-model.

Strong reshuffling of the original energy spectrum.

Energy landscape now depends very little on the sequence. The model is (maybe) very similar to

GD-model (Gaussian Disorder Model)

e_{ij}: Gaussian variables

$$\langle e \rangle = 0, \quad \langle e^2 \rangle = 1$$

and there is no reference to the basis sequence.

What do we do?

The ϵ-coupling method.

1. Compute a ground state structure (if the model is degenerate it will be one of the many GS structures).

$$\vec{l}^{(0)} \equiv \{l_{ij}^{(0)}\}$$

2. Modify the Hamiltonian H by adding a perturbation

$$\mathcal{H} \to \mathcal{H}' = \mathcal{H} - \epsilon(1 - q)$$

$$q = \frac{1}{L}\vec{l} \cdot \vec{l}^{(0)} = \frac{1}{L}\sum_{i,j} l_{ij} l_{ij}^{(0)}$$

   Notice that here we have defined q, the overlap with the "true GS", such that $q \geq 0$.

   This perturbation is a repulsive term in the space of structures.

3. Compute $\vec{l}^{\epsilon} \equiv \{l_{ij}^{\epsilon}\}$ under $\mathcal{H}'$.

4. Do the same things for many values of ϵ.

$\forall$ disorder set

$$d_J(\epsilon, L) \equiv 1 - \frac{1}{L} \vec{l}^{\epsilon} \vec{l}^{(0)}$$

and

$$\Delta E_J(\epsilon, L) \equiv E(\vec{l}^{\epsilon}) - E(\vec{l}^{(0)})$$

are non-decreasing functions of ϵ.

$\overline{\cdots}$ is the disorder average, over $\{r\}$ and $\{\eta\}$.

$$d(\epsilon, L) \equiv \overline{d_J(\epsilon, L)}$$

$$\Delta E(\epsilon, L) \equiv \overline{\Delta E_J(\epsilon, L)}$$

(The new GS $\vec{l}^{\epsilon}$ is found using exactly the same algorithm that one uses for $\vec{l}^{(0)}$: the ϵ-term is introduced as a modification of the original pairing energies e_{ij}).

An important quantity:

$$d_\infty(\epsilon) = \lim_{L \to \infty} d(\epsilon, L)$$

Are typical low energy structures all "very similar", or they can be "very different"?

$d_\epsilon = 0 \ \forall$ finite ϵ: "close".

$d_\epsilon > 0$ for $\epsilon > \epsilon^* \Longrightarrow$ "distant", i.e. structures with the same energy may have a broad probability distribution function of their distances and overlaps.

"Non trivial $P(q)$".

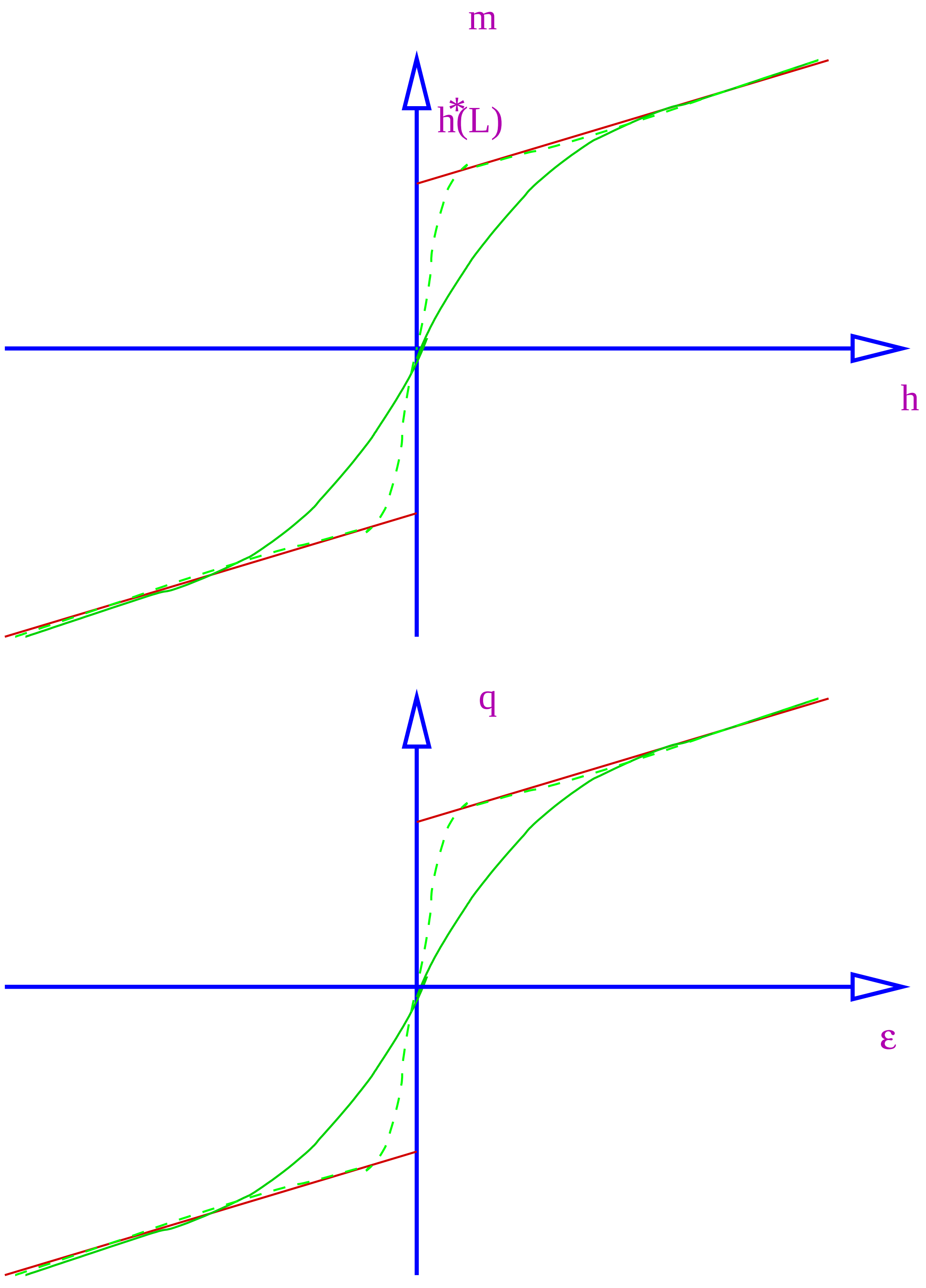
Assume that $\Delta E_{\text{typ}} \propto L^\theta \Longrightarrow$ $d(\epsilon, L) \propto \epsilon L^{-\theta}$ with the same θ (for dimensional reasons).

We can measure the same θ in two different ways: good check.

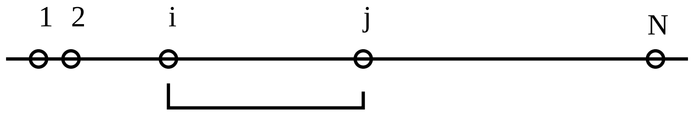
Other possible approach: pinch two basis (same for spin models: ϵ-coupling, PAP, fix 2 spins. What is typical?).

ϵ is equivalent of magnetic field for spin models

(check for symmetry breaking):



Let us look a moment in the method (simple, easy, effective...).
$Z_{i,j}$ partition function restricted to the sites between i and j.



$Z_{1,N} = Z$, $E = \sum_{(ij)\in\mathcal{S}} e_{ij}$ $Z = \sum_{\mathcal{S}} e^{-\beta E[\mathcal{S}]}$.

Recursion relation:

$$Z_{i,j} = Z_{i,j-1} + \sum_{k=i}^{j-1} Z_{i,k-1}\ e^{-\beta e_{k,j}} Z_{k+1,j-1}\ ,$$

$$Z_{i,i} = Z_{i,i-1} = 1\ \forall\ i$$

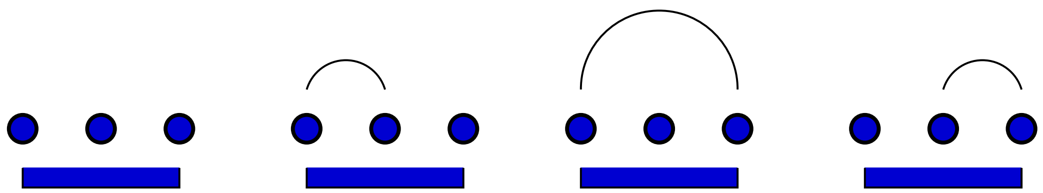
For example we start with two points:

$$Z_{1,2} = Z_{1,1} + Z_{1,0}\ e^{-\beta e_{1,2}} Z_{2,1} = 1 + e^{-\beta e_{1,2}}$$

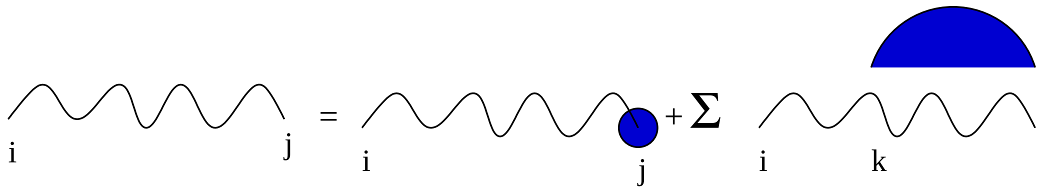
Let us go further and apply the recursion to a chain done of three basis:

$$Z_{1,3} = Z_{1,2} + Z_{1,0}\ e^{-\beta e_{1,3}} Z_{2,2} + Z_{1,1}\ e^{-\beta e_{2,3}} Z_{3,2}$$

$$= \left(1 + e^{-\beta e_{1,2}}\right) + e^{-\beta e_{1,3}} + e^{-\beta e_{2,3}}$$



The iteration rule is basically an Hartree-Fock scheme:



It is made possible by the planarity condition.

It is fast: $O(N^3)$ steps. Numerically up to $O(N) = 10^4$. Basis of the analytic approach.

The **D**(egenerate) model.

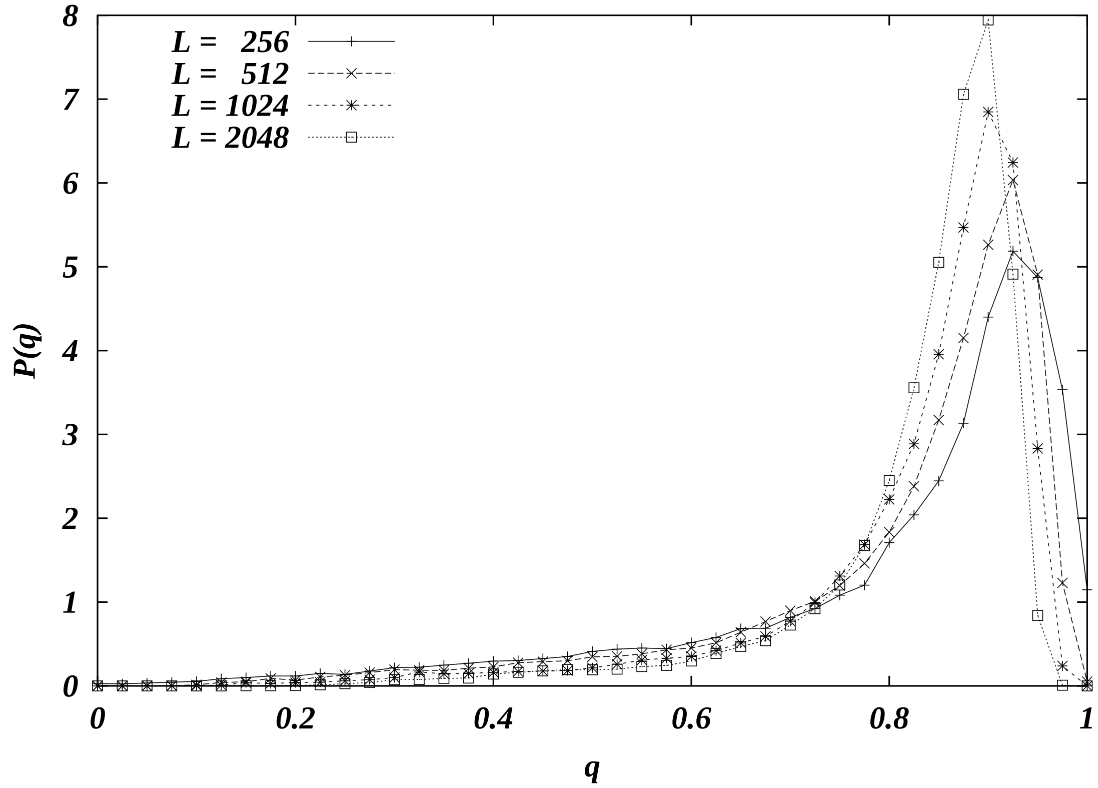
$T=0$ degenerate GS (exponentially large number): we can study $P(q)$, $q=\frac{1}{L}\sum_{ij} l_{ij}^{(0)} l_{ij}^{\epsilon}$.

We want to list them all!

But they are too many (large degeneracy: already at $L=256$ you cannot keep all the structures). So we add (only for this analysis) a further constraint.

We forbid structures where a single base pair is surrounded by non-paired basis ($l_{i-1,j+1}=l_{i+1,j-1}=0$ and $l_{ij}=0$ is forbidden). This is suggested by experimental observations.

Here: average over 100 samples.



Usual:

peak that moves;

decreasing width.

$$\text{peak position} \longrightarrow_{L\to\infty} q \simeq 0.87$$

$$\sigma^2 \sim L^{-0.45}$$

$$P(q) \longrightarrow_{L\to\infty} \delta(q-\overline{q})$$

Now the different low energy states.

Ground states for $\mathcal{H}'$,

for 18 values of $\epsilon \in (0.001, 131.072)$

equally spaced on log scale.

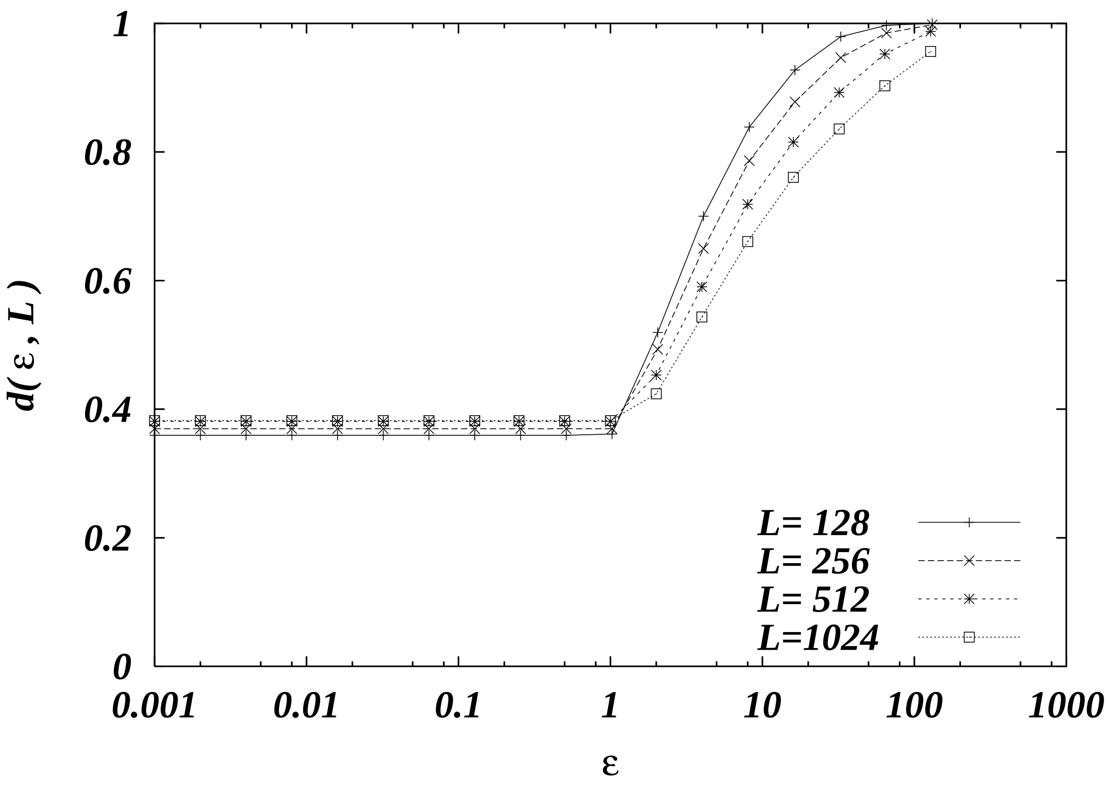
$L = 128$: 50000 samples...

$L = 4096$: 500 samples.

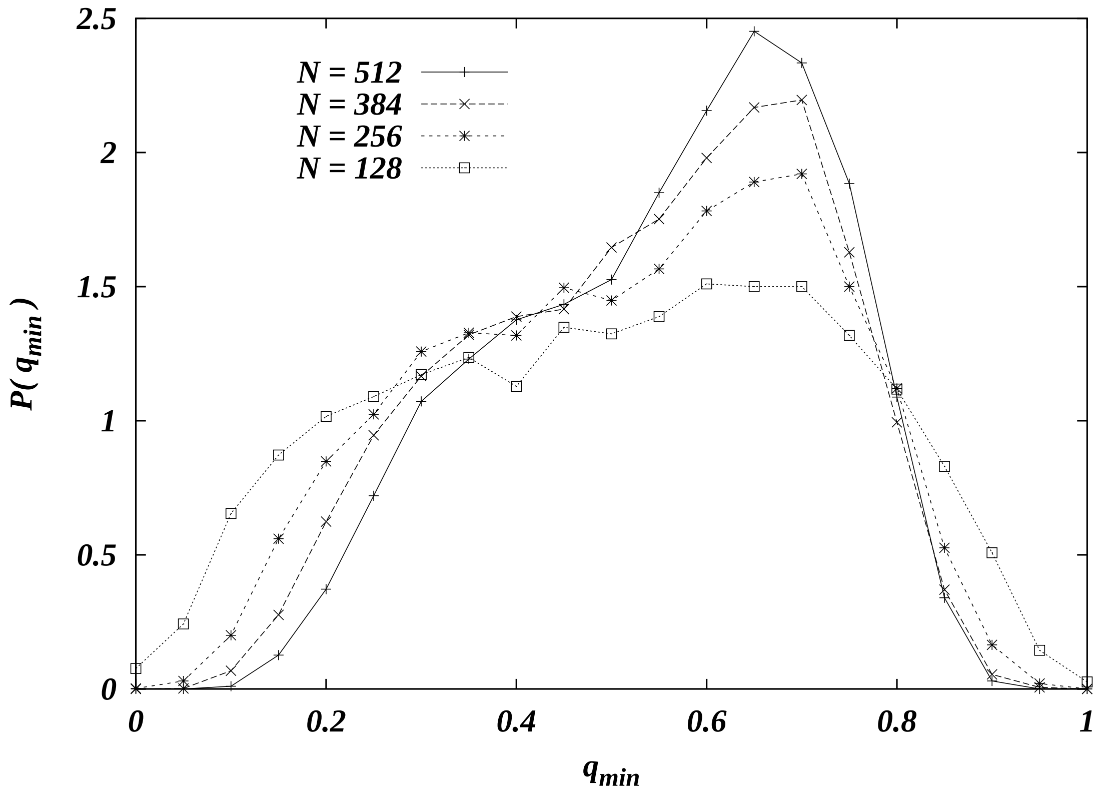
First $\vec{l}^{(0)}$ at random. Then switch ϵ on.

Nothing happens for $\epsilon < 1$ (since the perturbation is not strong enough to go out of the set of the degenerate ground states).

We are just selecting a different one among the original degenerate GS.



Plateau $d \sim 0.38 \Longrightarrow q = 1 - d \simeq 0.62$.
This is the radius of the sphere containing the set of all the GS.
We have seen that $P(q)$ is peaked at $q \simeq 0.87 \Longrightarrow$ GS are not uniform on the sphere, but they are dense in the central region and sparse on the boundary.
Two GS at random will typically be in the dense region, while forcing minimum overlap q_{min} (this is what we do in the ϵ-coupling method, where for given ΔE we have to go as far as possible) the result is different.
Distribution of q_{min} is broad.



The **QD** Model

Minimal remotion of the degeneracy.
Keep as much as possible of the sequence structure.

Two regimes:

$\epsilon > 1$: $\sim$ **D**-model;

$\epsilon \leq 1$: non-trivial behavior.

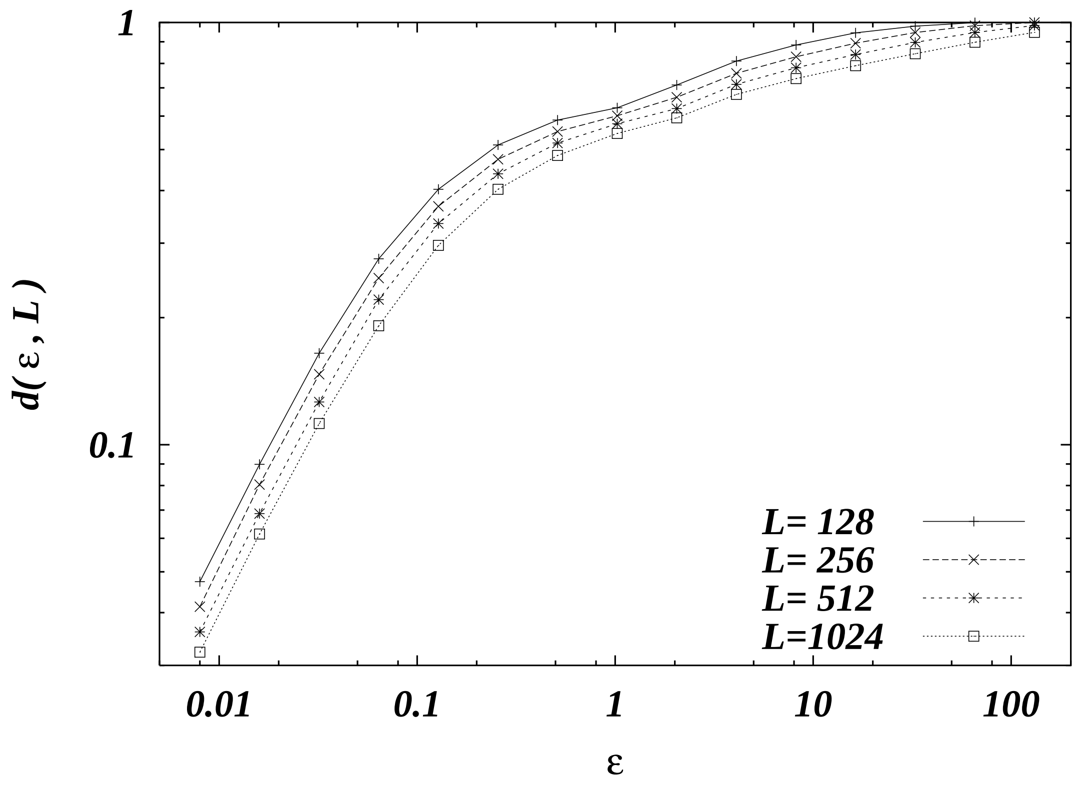
In the small ϵ region scaling behavior is not very clean, best fits are not decisive:

$\theta = 0$ is OK

$\theta \simeq 0.2$ is OK

$\log L$ scaling is OK

Probably like **ND**... see later

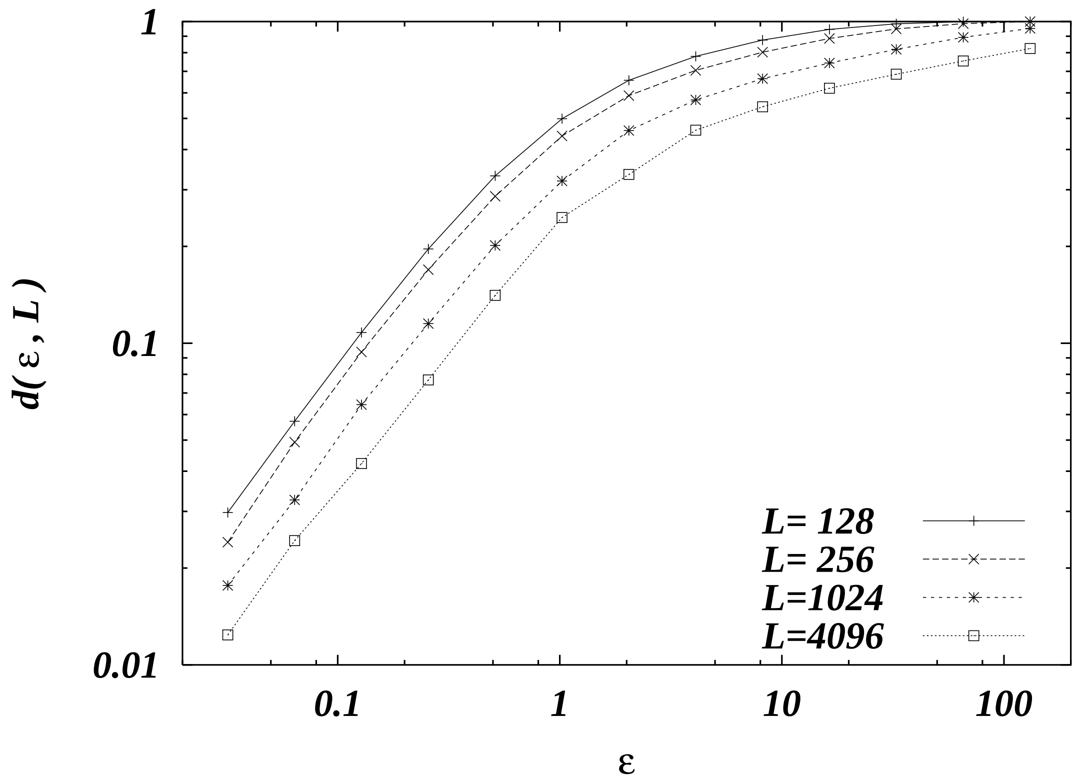


The **ND** Model

Here the GS degeneracy has been removed by a strong reshuffle of the energy levels.

We plot $d(\epsilon, L)$ versus ϵ for different L values.

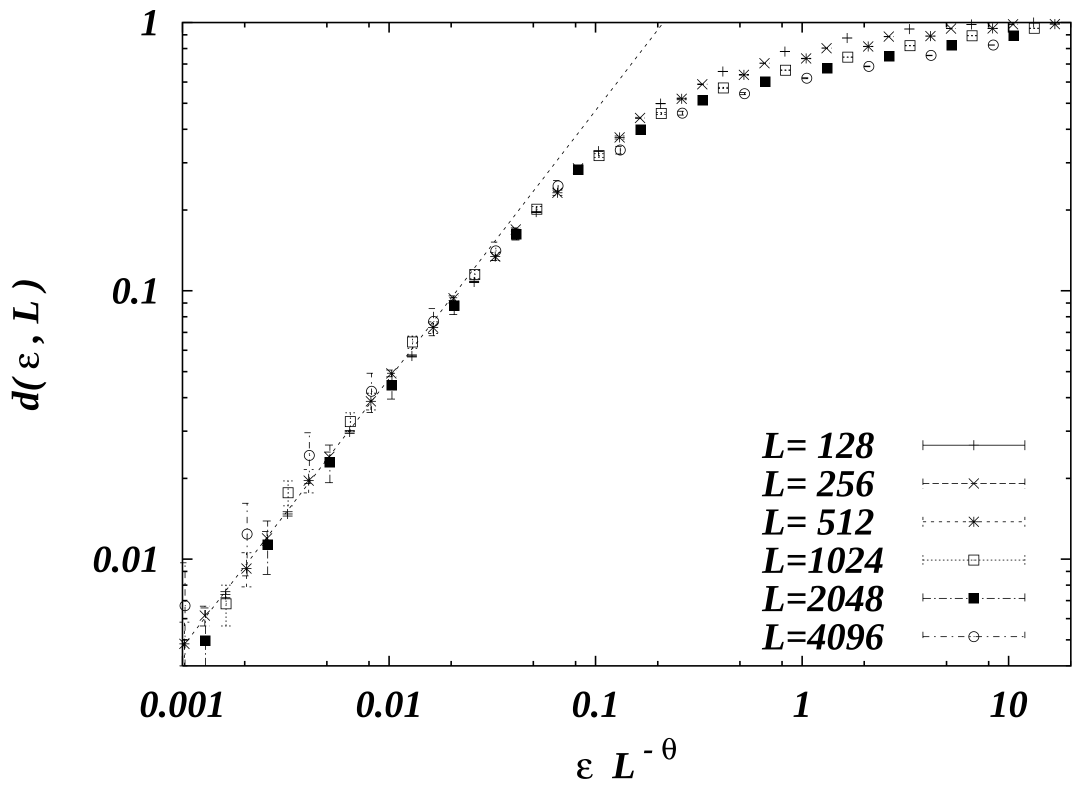
Now data are a smooth function of ϵ.



$d(\epsilon, L) \sim \epsilon L^{-\theta} \Longrightarrow$ plot d versus $\epsilon L^{-\theta}$.

Best fit: $\theta \sim \frac{1}{3}$

Scaling works well up to $d \sim 0.4 - 0.5$ (if perturbation is too large phase space changes).

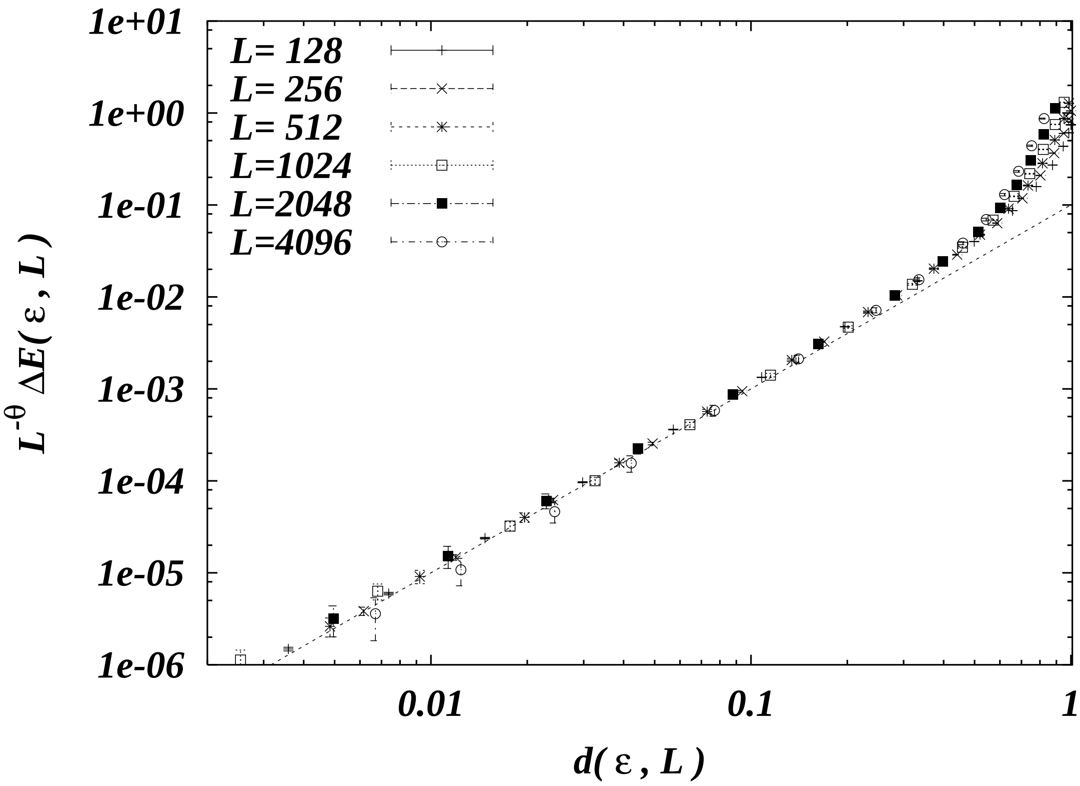


$\Delta E(\epsilon, L)$ versus $d(\epsilon, L)$ rescaled.

Best fit, again: $\theta \sim \frac{1}{3}$

Small d: $\Delta E(\epsilon, L) \propto d^2$.

$\theta = 0$ does not work (maybe $\Delta E \sim (\log L)^2$, but even this fit is not very good).



The **GD**-model (Gaussian Disorder)

Again:

$\longrightarrow$ $\theta \simeq \frac{1}{3}$ is best;

$\longrightarrow$ for small ϵ $d \simeq \epsilon$;

$\longrightarrow$ for small d, $\Delta E \propto d^2$;

$\longrightarrow$ results do not depend on details like the presence of the constraint $l_{ij} = 0$ for $|i - j| < 4$.

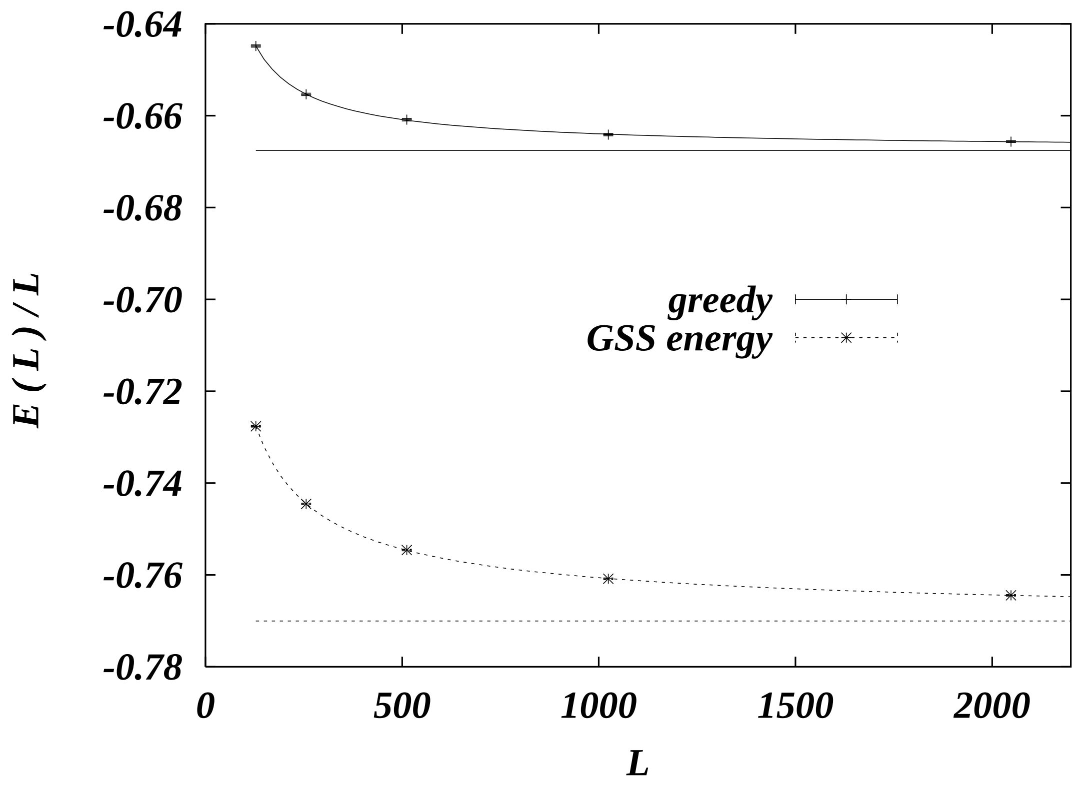
So, we have a very simplified model with the same behavior of the original (ND) one.

The model does not have a trivial energy landscape (horrible doubt...)

Idea to check: in a trivial energy landscape any good greedy algorithm should be able to get at least very close to the GS energy.

We try a greedy algorithm
start with an open structure
at each step link the lowest negative
base couple allowed by the
planarity constraint

We reach energies $O(10\%)$ higher than true GS energy.



- There is a glassy phase at low T.
- This phase is "Droplet"-like, i.e. $\theta > 0$.
- $\theta \sim \frac{1}{3}$
- Consistent results from different determinations.
- Different models are "very similar" (even if the quasi-degenerate model is difficult to qualify precisely).