RNA folding
and
Matrix Field Theory

Henri Orland (SPhT, Saclay)

Collaboration with

- A. Zee (KITP, UCSB)
- M. Pillsbury (UCSB)

    and

- G. Vernizzi (SPhT, Saclay)
- M. Bon (Saclay)

Outline

- Review of basic properties of RNA
- Secondary structures
- Matrix field theory for RNA
- Large N expansion
- Recursion relations
- Exact enumeration of RNA structures
- Topological classification of RNA
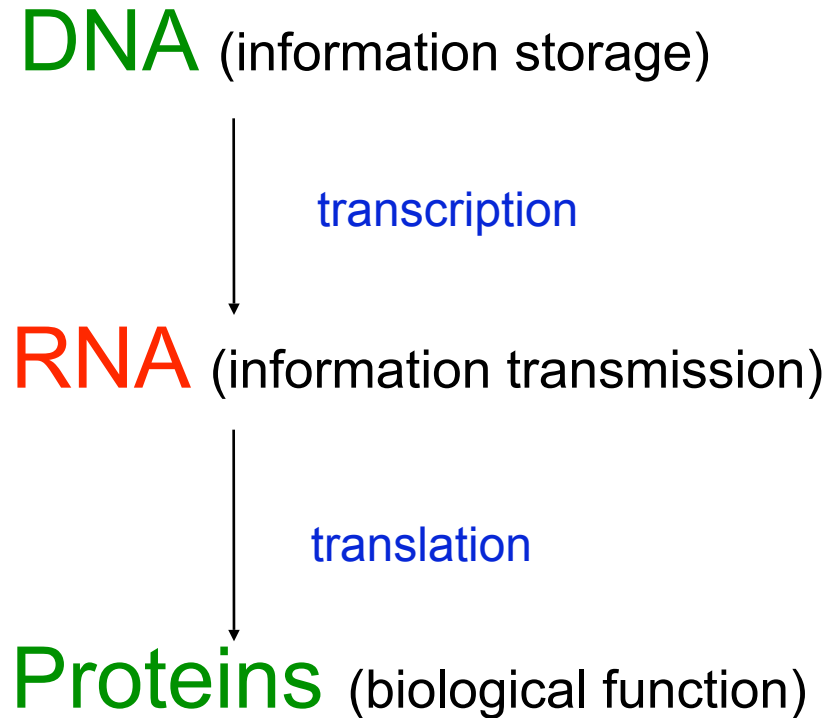- Monte Carlo approach

Review of basic properties of RNA

- RNA is a biopolymer
    - RNA (length ~ 70 – 2000)
    - DNA (length ~ 10^6 – 10^9)
    - Proteins (length ~ 10^2)
    - Polysaccharides (length ~ 10^3)

Composition of Cell (in weight)

- Water   70%
- Proteins   15%
- DNA   1%
- RNA   6%
- Polysaccharides   3%
- Lipids   2%
- Mineral ions   3%
- Etc…

Central dogma of Biology

DNA (information storage)

↓ *transcription*

RNA (information transmission)

↓ *translation*

Proteins (biological function)

Several forms of RNA

- Messenger : mRNA (L ~ 1000)
- Transfer: tRNA (L ~ 70)
- Ribosomal: rRNA (L ~ 3000)
- Micro: μRNA (L ~ 25)
- Huge amounts of non-coding RNA in "junk" DNA

Why does the 3d structure of RNA matter?

Important discovery in the 80s: RNA can have enzymatic activity

Important discovery since 2000: µRNA play crucial role in cell regulation

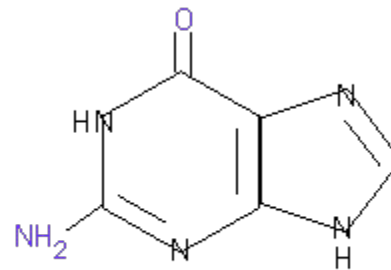Function strongly related to shape

⟶ Must know 3d structure of RNA

Chemistry of RNA

- RNA is a heteropolymer

- Four bases:
  - Adenine (A)
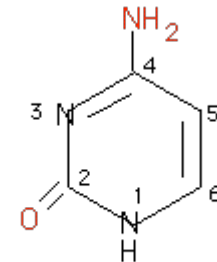 - Guanine (G)
  - Cytosine (C)
  - Uracil (U)

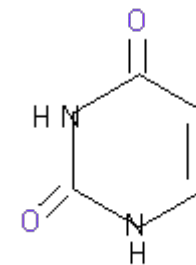 The sugar phosphate backbone polymerizes into a single stranded charged (-) polymer
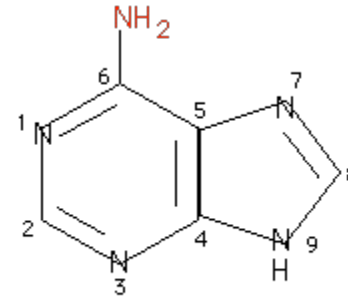
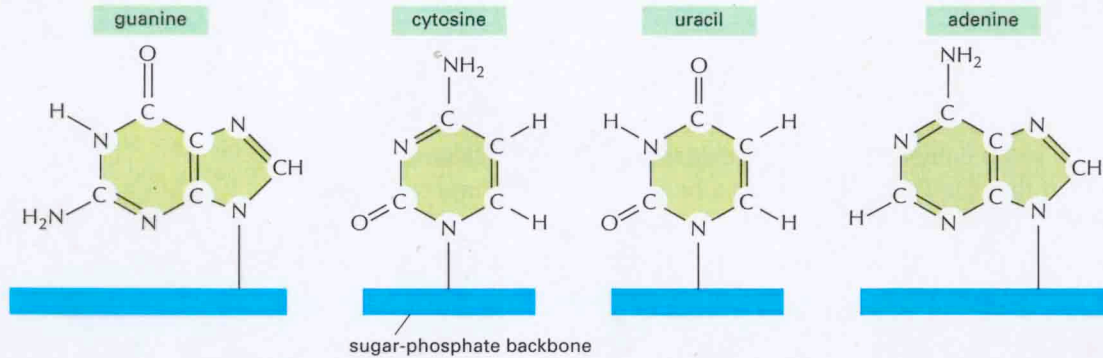FOUR BASES OF RNA

guanine   cytosine   uracil   adenine

sugar-phosphate backbone

SUGAR-PHOSPHATE BACKBONE OF RNA

base

ribose

5' end

3' linkage   5' linkage

phosphodiester bond

3' end
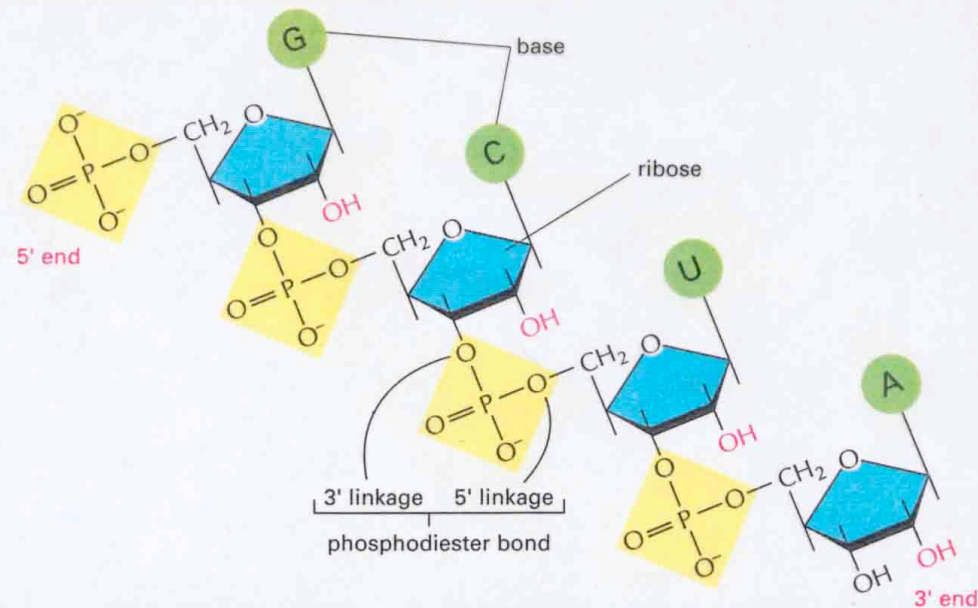
Energy scales

- Crick-Watson: conjugate pairs

$$C - G$$
$$A - U$$

Pairing due to Hydrogen bonds between bases $\Longrightarrow$ RNA folding

Stacking of aromatic groups

Electrostatics (Mg^{++} ions) controls 3d structure

Energy scales
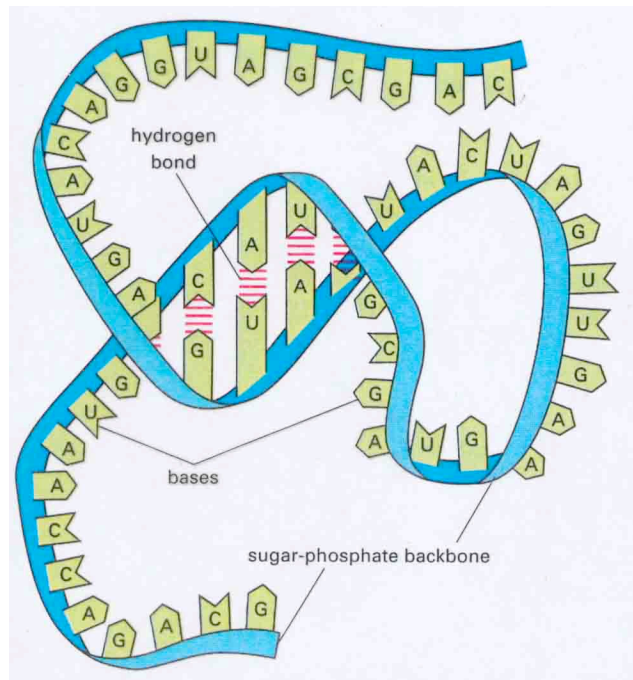
C — G : 3kCal/mole = 5 kT

A — U : 2kCal/mole = 3.3 kT

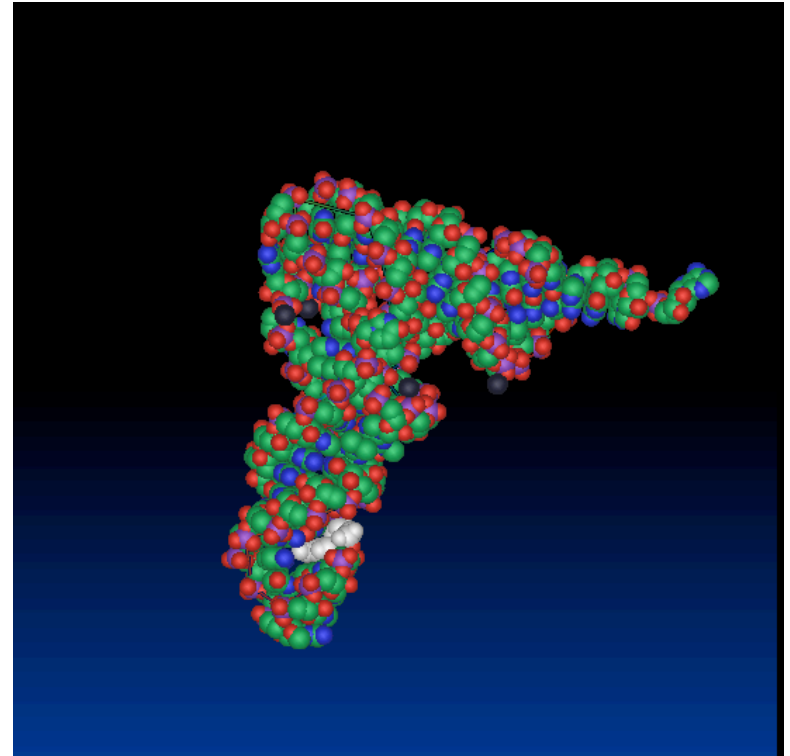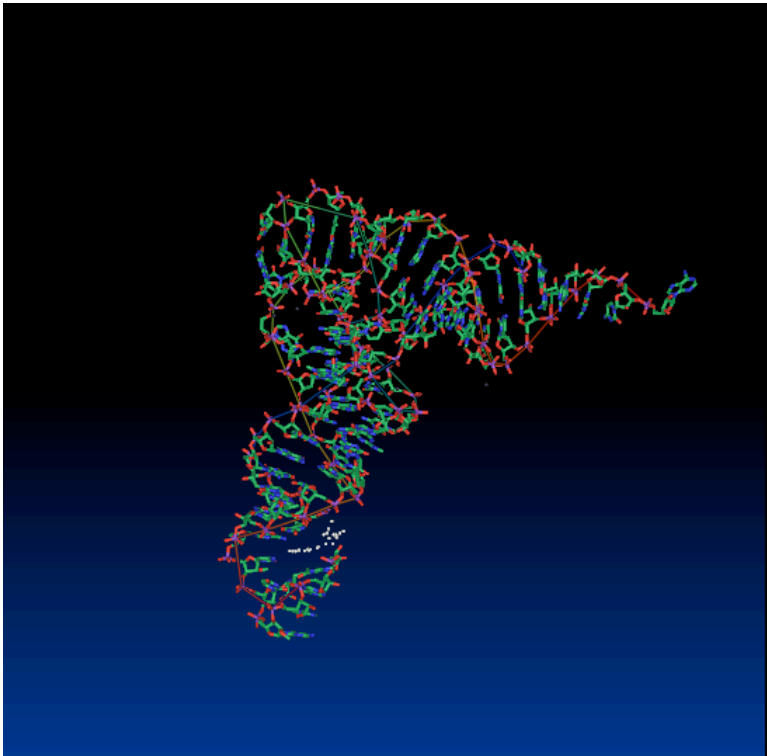G — U : 1kCal/mole =1.6 kT

300 K = 0.6 kCal/mole = 1/40 eV

Base pairing

- Induces helical strands (like in DNA)
- Induces secondary structure of RNA

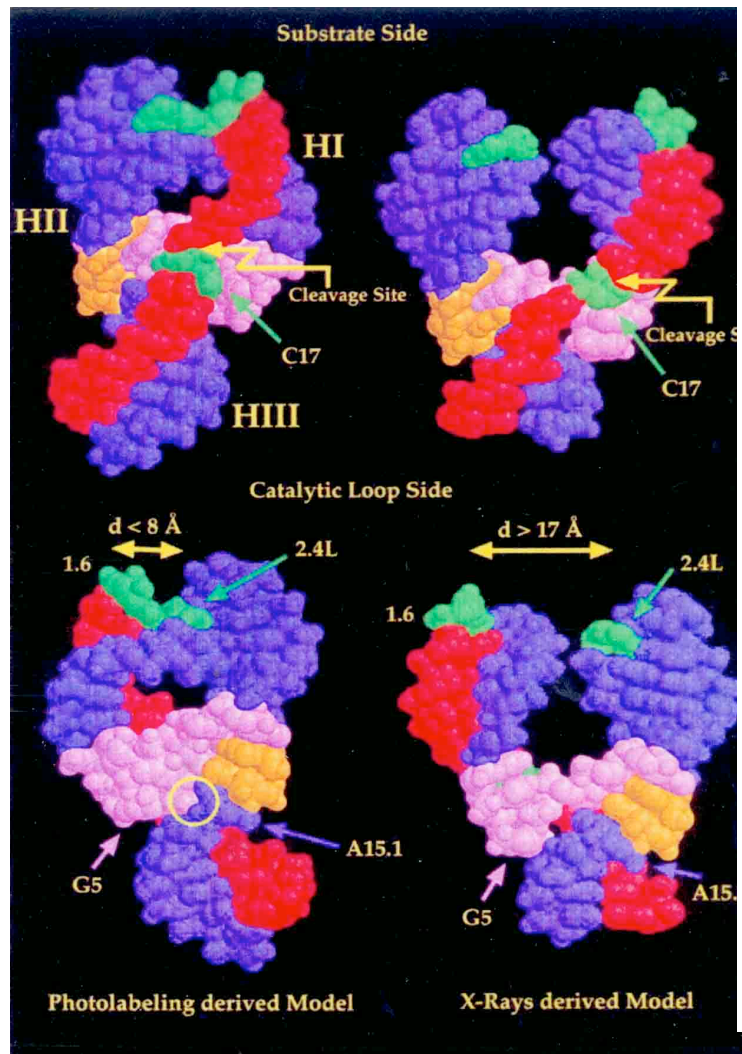# Pictures of RNA

Transfer RNA

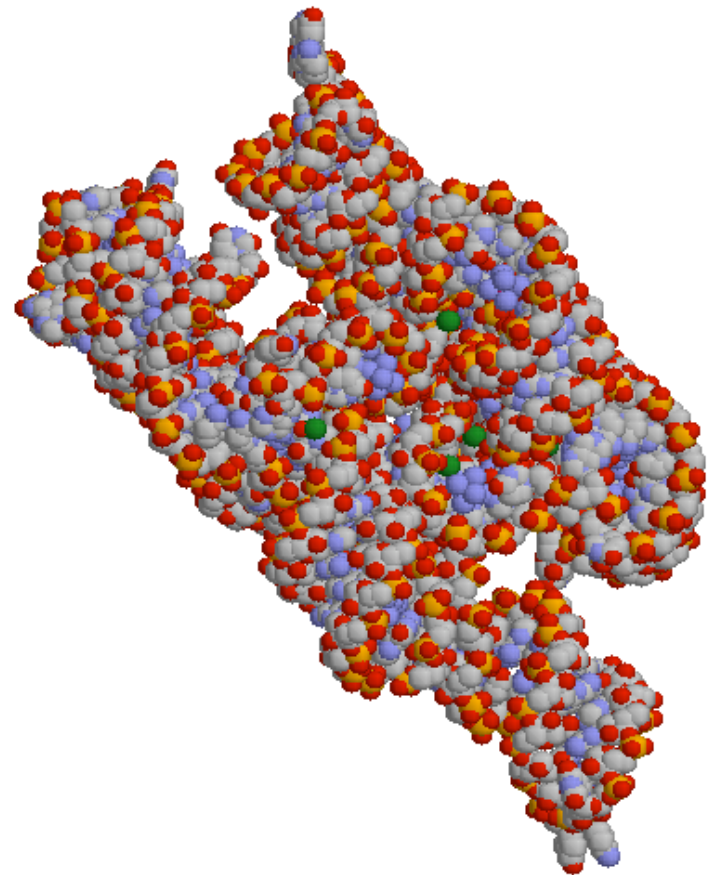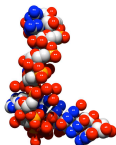H. Orland, SPhT, Saclay   RNA folding, Trieste 2006

Hammerhead Ribozyme
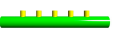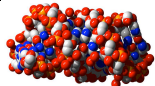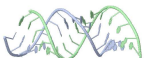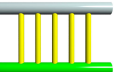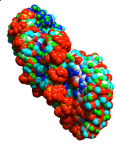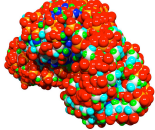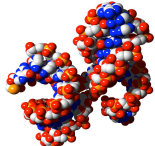
Ribosomal RNA

Secondary structures

- In RNA, there are helical stems with loops and bulges

Spacefill view	3D structure	Secondary structure

Pseudo-knots in RNA

- In addition to secondary structure, there are "pseudo-knots" which constrain the 3d structure

the Kissing Hairpin

The H pseudo-knot

Loop-bulge
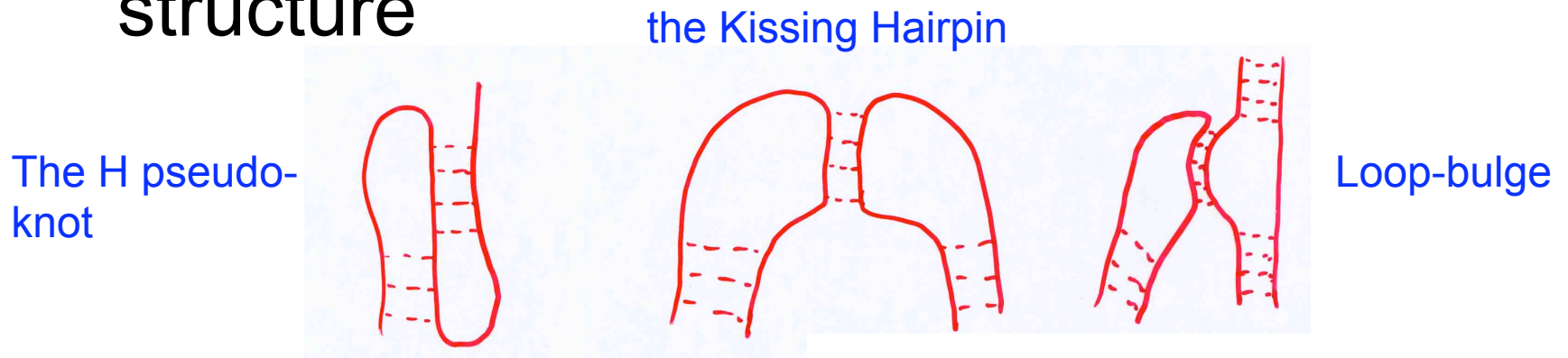
- 3d folding controlled by concentration of Mg^{++} ions.

In fact base pairing is not good enough: need also stacking energies.

However:
- saturation of Crick-Watson pairing
- pseudo-knot free energy << free energy of secondary structure

⟹ RNA folding much easier than protein folding

Partition function
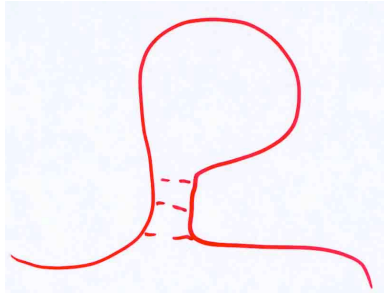
$v_{ij}(\vec{r}_{ij})$: interaction of base *i* and *j*

- short range
- saturating
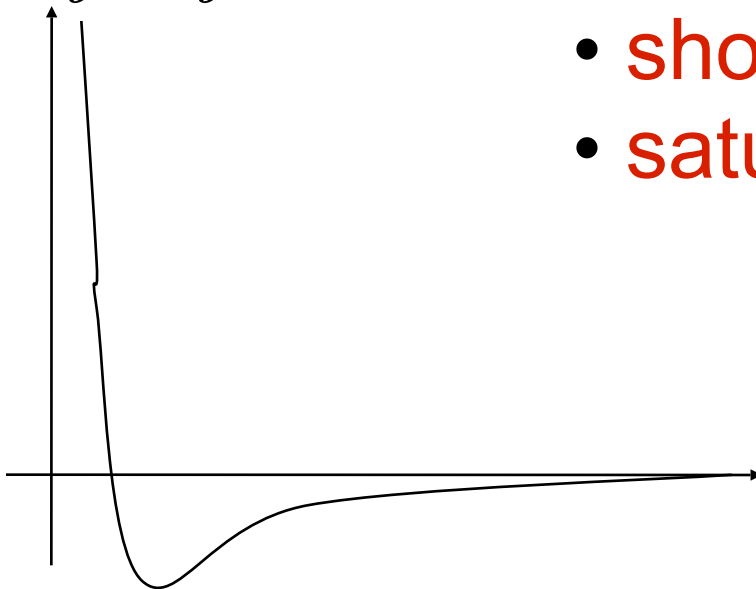
Partition function

$$Z = \int \prod_{i=1}^{L} d^3 r_i \prod_{i=1}^{L-1} \underbrace{f(\vec{r}_{i+1}, \vec{r}_i)} Q(\{\vec{r}_i\})$$

Interactions

Chain
connectivity

$$\longrightarrow \begin{cases} \delta(|\vec{r}_{i+1} - \vec{r}_i| - a) \\ e^{-\frac{3}{2a^2}(\vec{r}_{i+1} - \vec{r}_i)^2} \end{cases}$$

$$Q = e^{-\frac{\beta}{2} \sum_{i \neq j} v_{ij}(\vec{r}_{ij})} + \text{solvent} + \text{electrostatics}$$

Further simplifications:

– Saturation of interactions

– Watson-Crick pairing

Define

$$V_{ij} = e^{-\beta \varepsilon_{ij}} \theta(|i - j| - 4)$$

Base pair energy

Chain rigidity

• Approximation

$$Z = \sum Q_0$$

sterically allowed configurations

where

$$Q_0 = 1 + \sum_{i<j} V_{ij} + \sum_{i<j<k<l} (V_{ij}V_{kl} + V_{ik}V_{jl} + V_{il}V_{jk})$$

$$+ \ldots + \sum_{i<j<k<l<\ldots<p<q} V_{ij}V_{kl} \ldots V_{pq}$$

- sum is mainly combinatorial
- any index appears once and only once (saturation)

- In using this partition function, we have not taken into account the entropy of loops.
- For a loop of size *l*, the entropy is

$$S = l \log \mu - c \log l$$

- In fact the $\log \mu$ goes into the free energies of pairing, so that

$$S = -c \log l$$

- with $c = 3/2$ (Gaussian chain)
- $c = 1.75$ (Self Avoiding Walk)

Secondary structures

- We work on Q_0

- Secondary structures = Arches
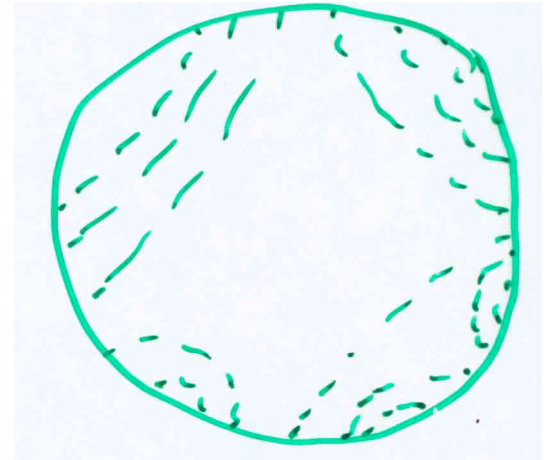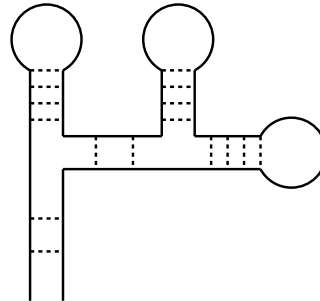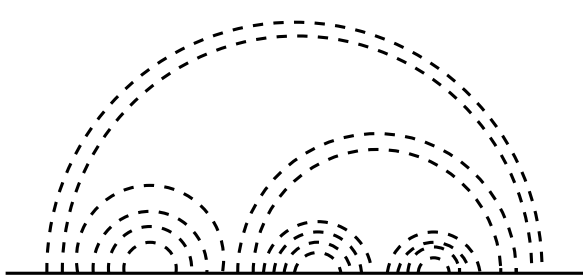
- Define $Z(i,j)$ as the

- partition function of segment (i,j)

Recursion relation

- Graphically, when one adds a base
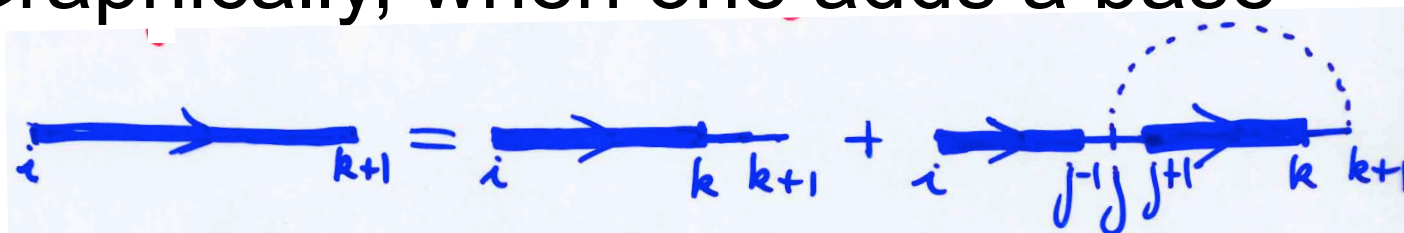
$$Z(i, k+1) = Z(i, k) + \sum_{j=1}^{k} V_{j,k+1} Z(i, j-1) Z(j+1, k)$$

- with

$$V(i, j) = e^{-\beta \varepsilon(i,j)} \theta(|i - j| - 4)$$

- by iterating this recursion, one can generate all possible secondary structures, with correct Boltzmann weights.

- This is the best tool for predicting secondary structures in RNA : more than 85% of base pairings correctly predicted

- Algorithm scales as N^3

- One can include Entropies and Stacking Energies

- http://www.tbi.univie.ac.at/~ivo/RNA Vienna Package

Bacillus subtilis RNase P RNA

M - multi-loop

I - interior loop

B - bulge loop

H - hairpin loop
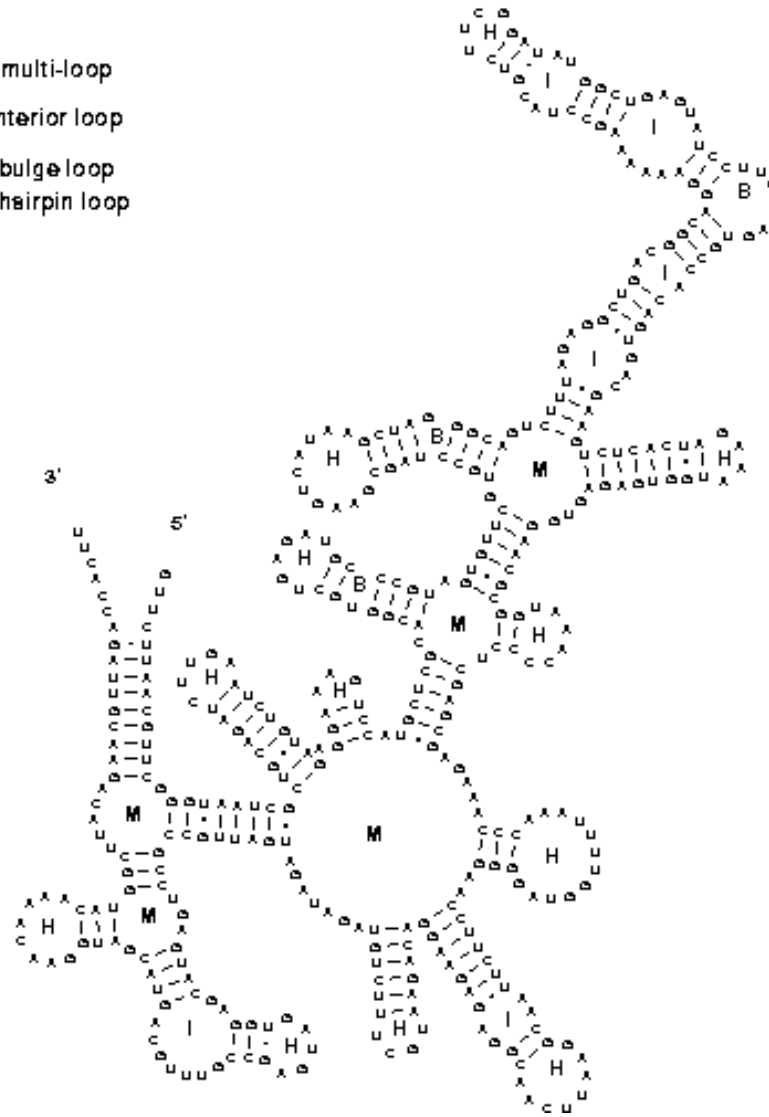
- Recursion equation looks like Hartree equations (tree diagrams)
- No Pseudo-Knots
- Is it possible to find a field theory such that secondary structures are the Hartree graphs?
- Then, Pseudo-Knots would appear as the corrections to Hartree approximation.

Matrix Field Theory
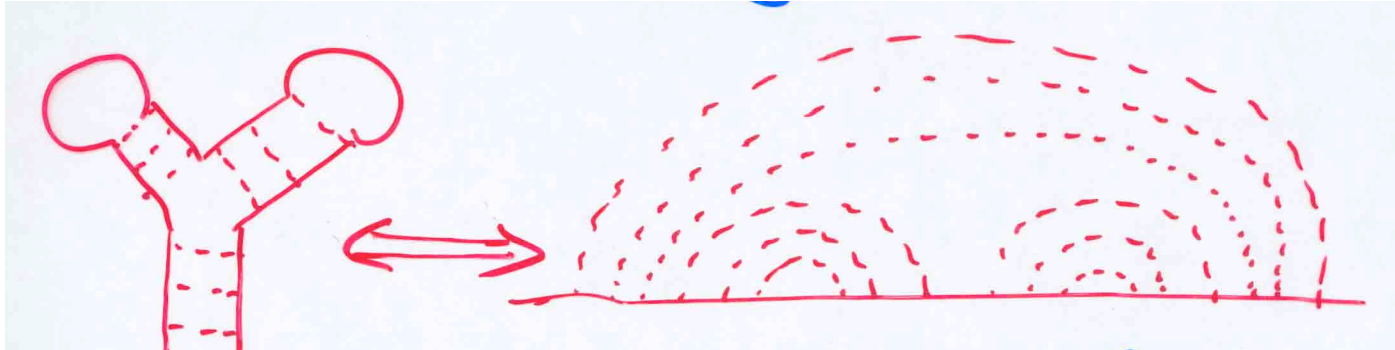
$$Q_0 = 1 + \sum_{i<j} V_{ij} + \sum_{i<j<k<l} (V_{ij}V_{kl} + V_{ik}V_{jl} + V_{il}V_{jk})$$

$$+ \ldots + \sum_{i<j<k<l<\ldots<p<q} V_{ij}V_{kl}\ldots V_{pq}$$

Wick Theorem

- Simple representation: consider an RNA sequence of length *L*

$$Q_0 = \frac{1}{\mathcal{N}} \int \prod_{i=1}^{L} d\phi_i \, e^{-\frac{1}{2} \sum_{i,j} \phi_i V_{ij}^{-1} \phi_j} \prod_{i=1}^{L} (1 + \phi_i)$$

- due to Wick theorem

$$V_{ij} = \frac{1}{\mathcal{N}} \int \prod_{i=1}^{L} d\phi_i \, e^{-\frac{1}{2} \sum_{i,j} \phi_i V_{ij}^{-1} \phi_j} \, \phi_i \phi_j$$

Wick Theorem

$$V_{ij}V_{kl} + V_{ik}V_{jl} + V_{il}V_{jk} = \frac{1}{\mathcal{N}} \int \prod_{i=1}^{L} d\phi_i \, e^{-\frac{1}{2} \sum_{i,j} \phi_i V_{ij}^{-1} \phi_j} \, \phi_i \phi_j \phi_k \phi_l$$

- However, this form gives same weight to all pairings. No penalty for Pseudo-Knots.
- Experimentally, few pseudo-knots.

- We look for a parameter *N* such that

$$N \rightarrow +\infty \equiv \quad \text{Secondary structures}$$

- $\quad$ Corrections in $\dfrac{1}{N} \equiv$ Pseudo-Knots

- Pseudo-knots are tunable by $[Mg^{++}]$ concentration

$$\dfrac{1}{N} \quad \text{plays the role of} \quad [Mg^{++}]$$
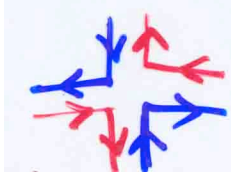
- **TOPOLOGY=MATRIX FIELD THEORY**

Matrix Field Theory: a Short Tutorial

- Vector field theories: $O(n)$ models count number of connected component of a graph. n is the fugacity of a loop.

- Matrix field theories: "count" topology.

- Consider the generalization of the scalar ϕ^4 field theory (t'Hooft, 1973)

- Consider the fields $\phi_{ab}(x)$: a $N \times N$ matrix at each point x in space.
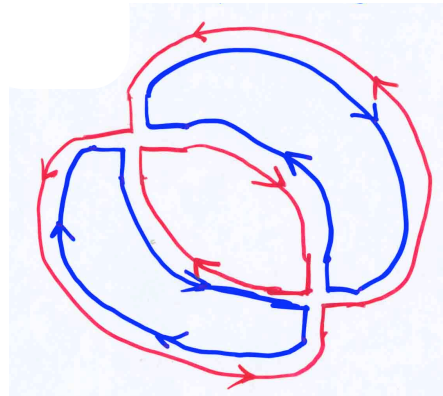
Matrix Field Theory

- A matrix ϕ^4 field theory is defined by

$$Z = \int \mathcal{D}\phi_{ab}(x) e^{-\frac{N}{2}\int dx Tr\phi(x)(-\nabla^2+m^2)\phi(x) - \frac{gN}{4!}\int dx Tr\phi^4(x)}$$

- represent $\phi_{ab}(x)$ by a double line

- Vertex: $NTr\phi_{ab}^4(x) \implies$ factor N

- Propagator: $\dfrac{1}{N}G(x-y) \implies$ $\dfrac{1}{N}$

Feynmann Graphs

- *V:* vertices
- *I:* internal propagators $\qquad \Longrightarrow \quad N^{V-I+L}$
- *L:* loops

- *V=2*
- *I=4*
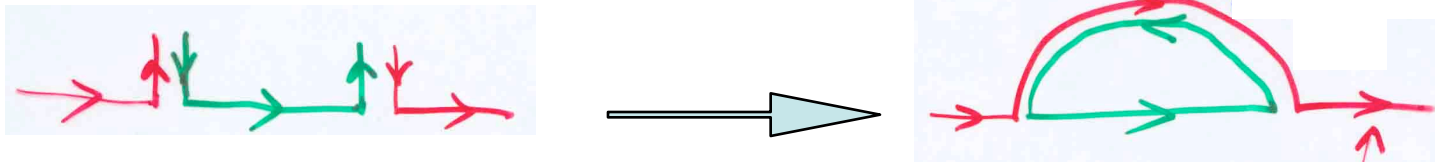- *L=4*
- **Euler characteristic:** $\quad \chi = V - I + L$

Euler characteristic and the Genus

- Consider a graph with Euler characteristics χ

- Theorem: this graph can be drawn without crossings on a surface of genus g given by $g = \dfrac{2 - \chi - c}{2}$ where c is the number of boundaries of the graph

- The genus g is the number of handles of the embedding surface

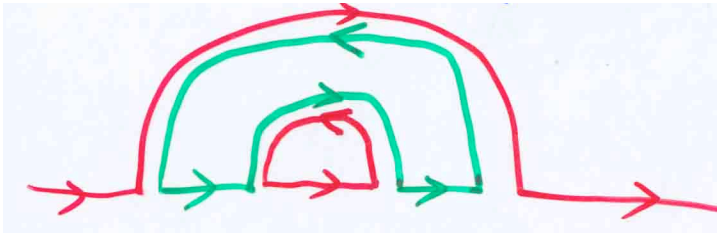# Double line graphs

- In our problem, if we use matrix fields

$\phi_{ab}(x)$: *NxN* matrix

- If we use same rule:

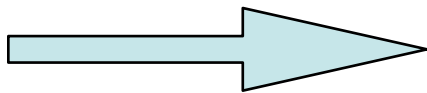Propagator: *1/N*
Loop: *N*

- Above graph: $N \times \dfrac{1}{N} = 1$
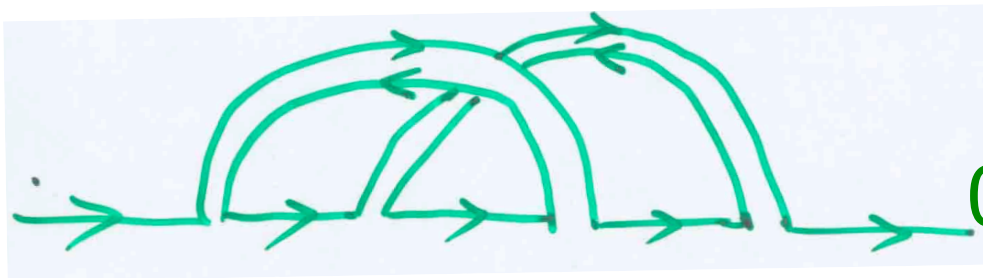
- Other graph

2 internal lines: $1/N^2$

2 Loops: N^2

  Order 1

- **Arches are of order 1**
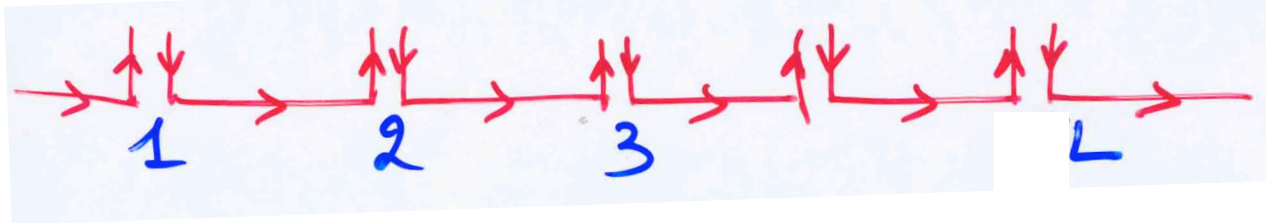

2 internal lines: $1/N^2$

0 Loops: 1

- **Pseudo-knots are of higher order in *1/N***

Matrix field representation of RNA folding

- We thus generalize the Wick theorem
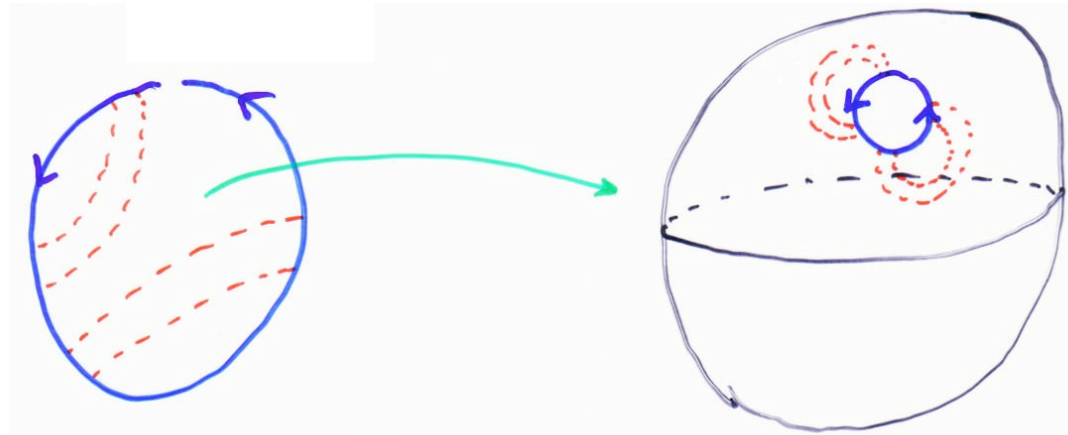
$$Z(1, L) = \frac{1}{A(L)} \int \prod_{k=1}^{L} d\varphi_k e^{-\frac{N}{2} \sum_{ij} (V^{-1})_{ij} \operatorname{tr}(\varphi_i \varphi_j)} \frac{1}{N} \operatorname{tr} \prod_{1}^{L} (1 + \varphi_l)$$

- By looking at a few diagrams, it seems to do what we want: Hartree diagrams are of order 1, pseudo-knots are of higher order.

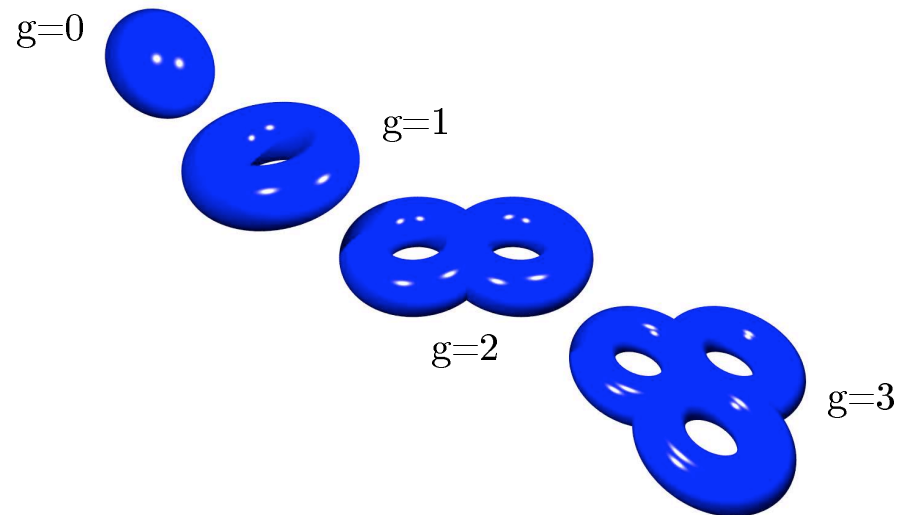# Topological classification of RNA folds
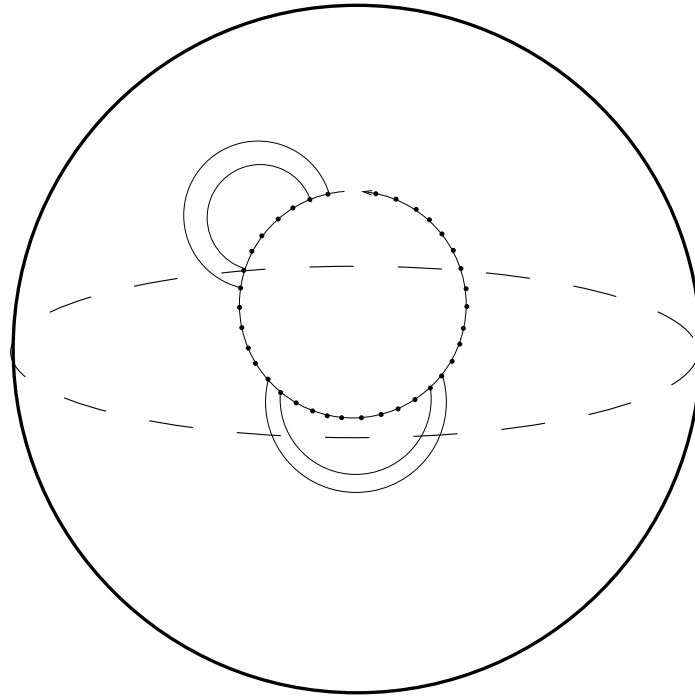
- An RNA fold can be characterized by its topology:

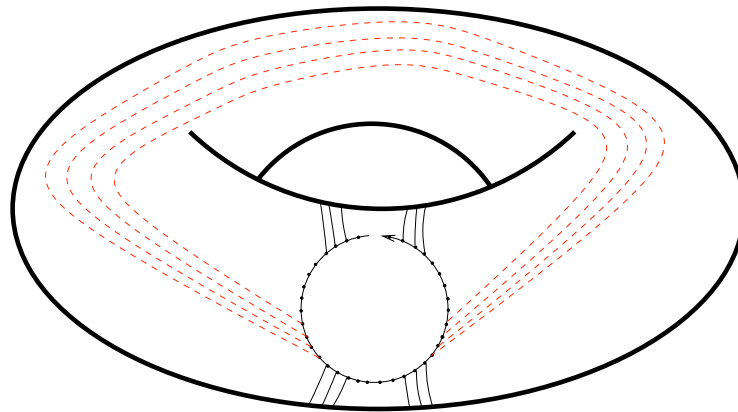- Number of handles of embedding surface

$$g = \frac{P - L}{2}$$
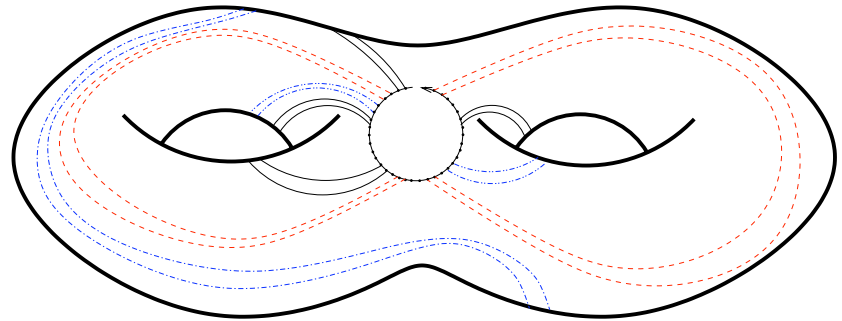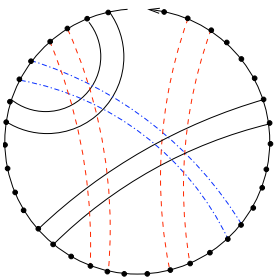
Topological expansion of closed oriented surfaces

g=0

g=1

g=2

g=3

Genus 0: the Sphere

Genus 1: the Torus

Genus 2: the Bi-torus
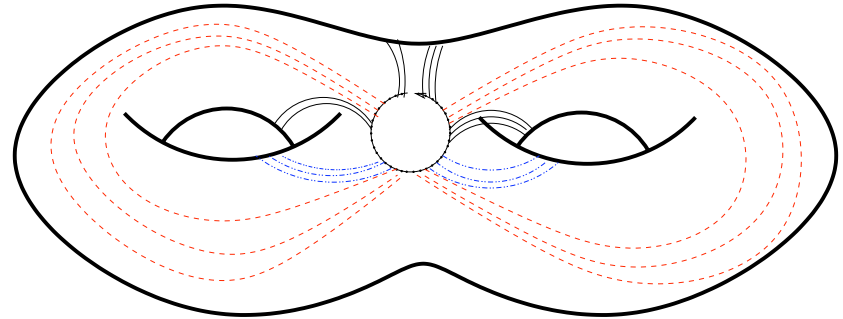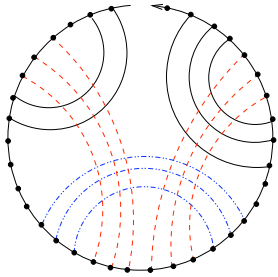
Genus 3
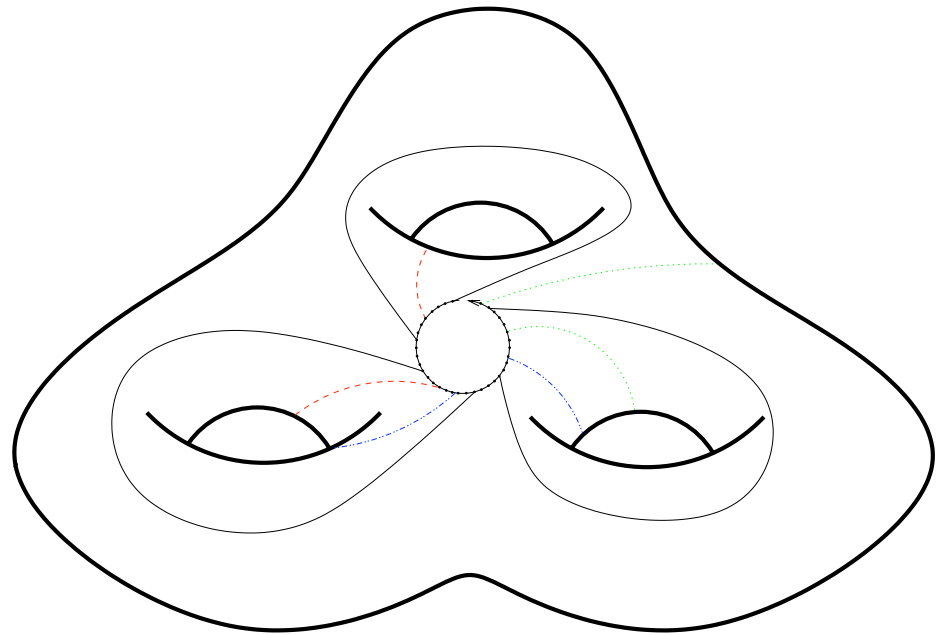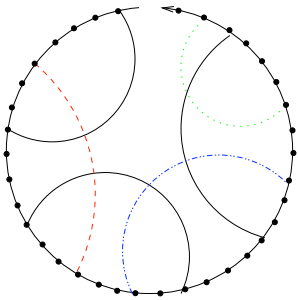
Large *N* expansion

- After some algebraic manipulations, one has the exact expression:

$$Z(1, L) = \frac{1}{C} \int dA \, e^{-\frac{N}{2} \operatorname{tr} A^2 + N \operatorname{tr} \log M(A)} M^{-1}(A)_{L+1,1}$$

- where $A_{ll'}$ is a $L \times L$ matrix and

$$M_{ij} = \delta_{ij} - \delta_{i,j+1} + i(V_{i-1,j})^{\frac{1}{2}} A_{i-1,j}$$

- The N dependence is explicit
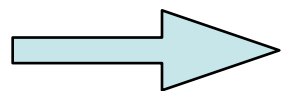
$\Longrightarrow$ one can perform a loop expansion (saddle-point)

The loop expansion

- Saddle-point equation

$$\frac{\partial S}{\partial A_{ll'}} = 0 \quad \Longleftrightarrow \quad \text{Hartree recursion equations}$$

- Expansion in *1/N*

$$A_{ll'} = A_{ll'}^{(0)} + \frac{x_{ll'}}{\sqrt{N}}$$

- Propagators of $x_{ll'}$ satisfy a Bethe-Salpeter equation

Bethe-Salpeter equation

- ## No order *1/N* correction
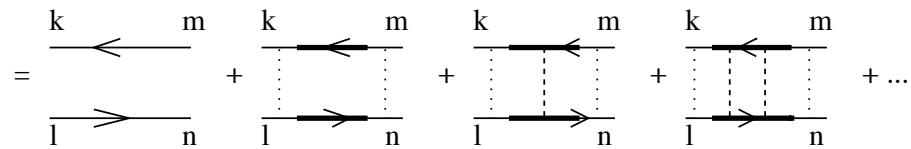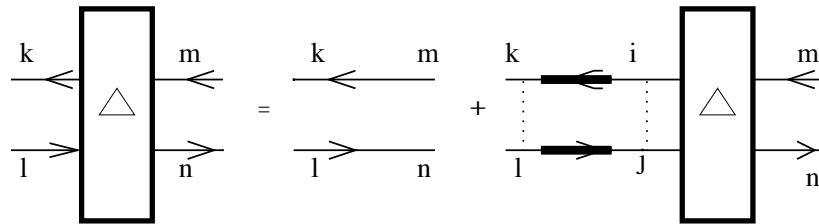
Parallel pairings don't change the genus
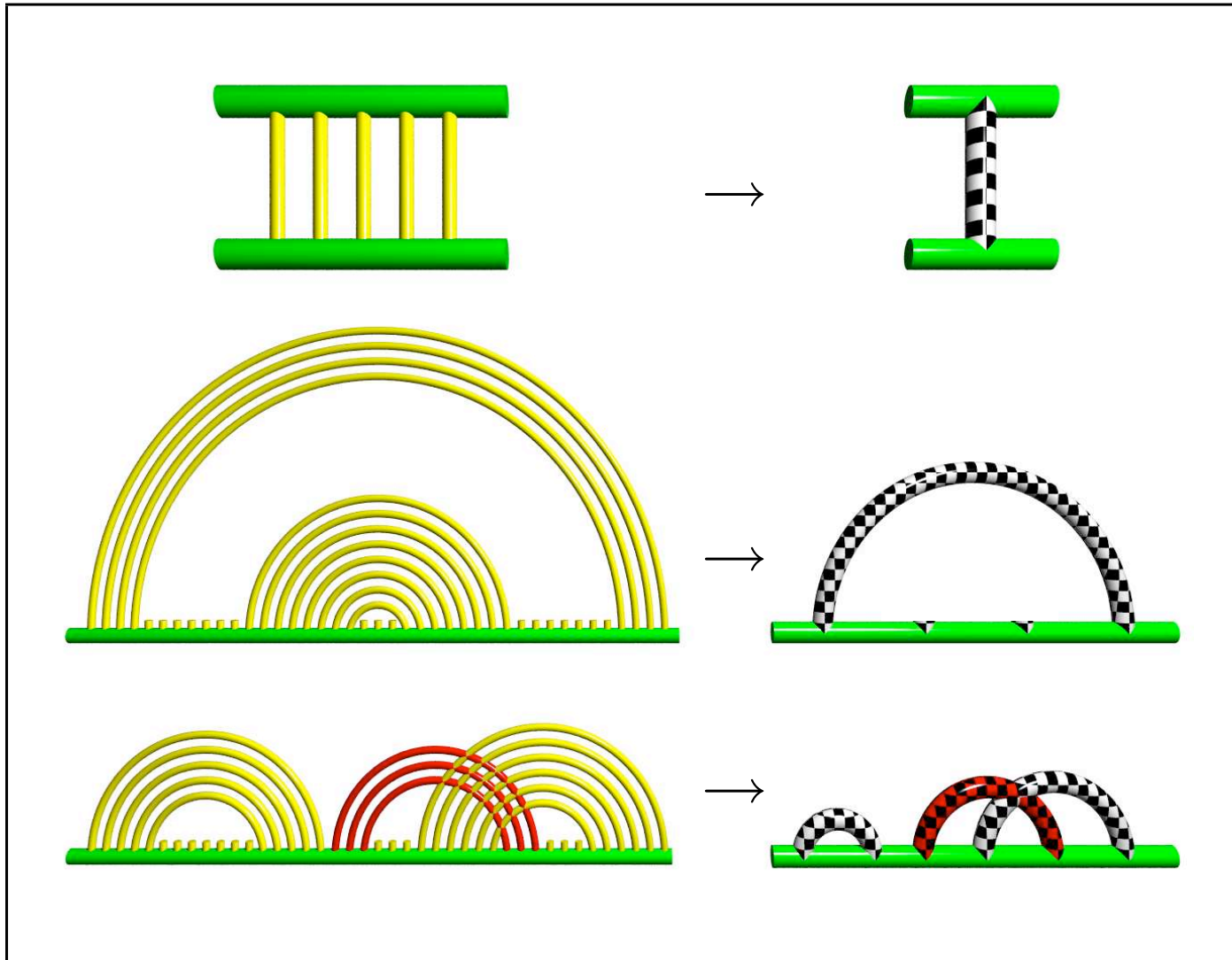
Recursion relations

- It is possible to obtain exact recursion relations for genus 1

- There is an exact relation

$$Z(1, L+1) = Z(1, L) + \sum_{k=1}^{L} V_{L+1,k} < \frac{1}{N} Tr \prod_{i=1}^{k-1}(1+\phi_i) \times \frac{1}{N} Tr \prod_{j=k+1}^{L}(1+\phi_j) >$$

- which can be expanded in powers of $\frac{1}{N}$

- Algorithm scales as L^6 $\Longrightarrow$ too long!

Irreducibility and Nesting
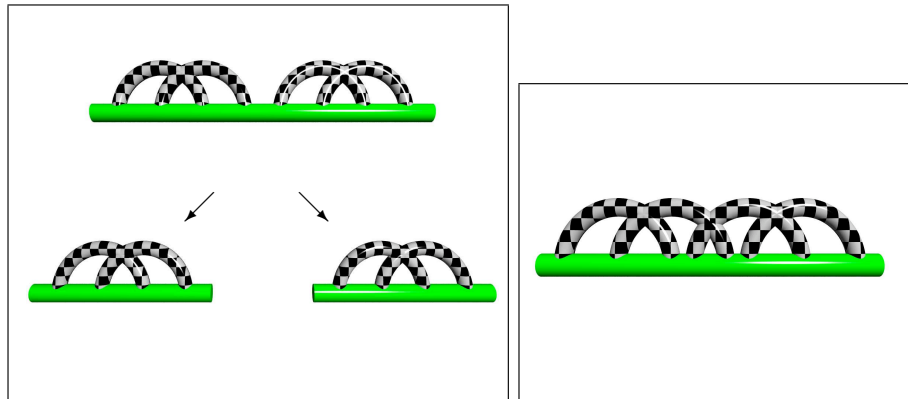
Irreducible PK

Genus is additive
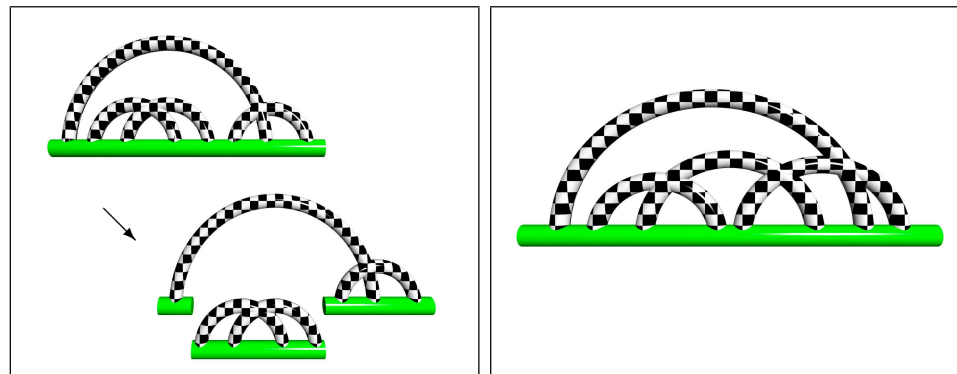
Non nested PK

Primitive Pseudo-Knots

Irreducible and
non-nested
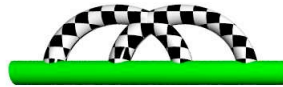
Only 4
primitive PK
of genus 1

Statistical study

- Look in database and calculate genii of pseudo-knots
- PseudoBase: around 245 pseudo-knots; all are of genus 1, except 1 of genus 2
- 237 H PK of the type ABAB
- 6 KHP of the type ABACBC
- 1 PK of the type ABCABC
- 1 PK of type ABCDCADB with genus 2

- Protein Data Bank (PDB): 850 RNA Structures

- 650 RNA have genus 0 (short fragments)

- Number of bases ranges from 22 (H PK with genus 1) to 2999 (with genus 15)

- Maximum total genus is 18. Maximum genus of primitive PK is 8.

- Transfer RNA (L=78) are KHP of genus 1
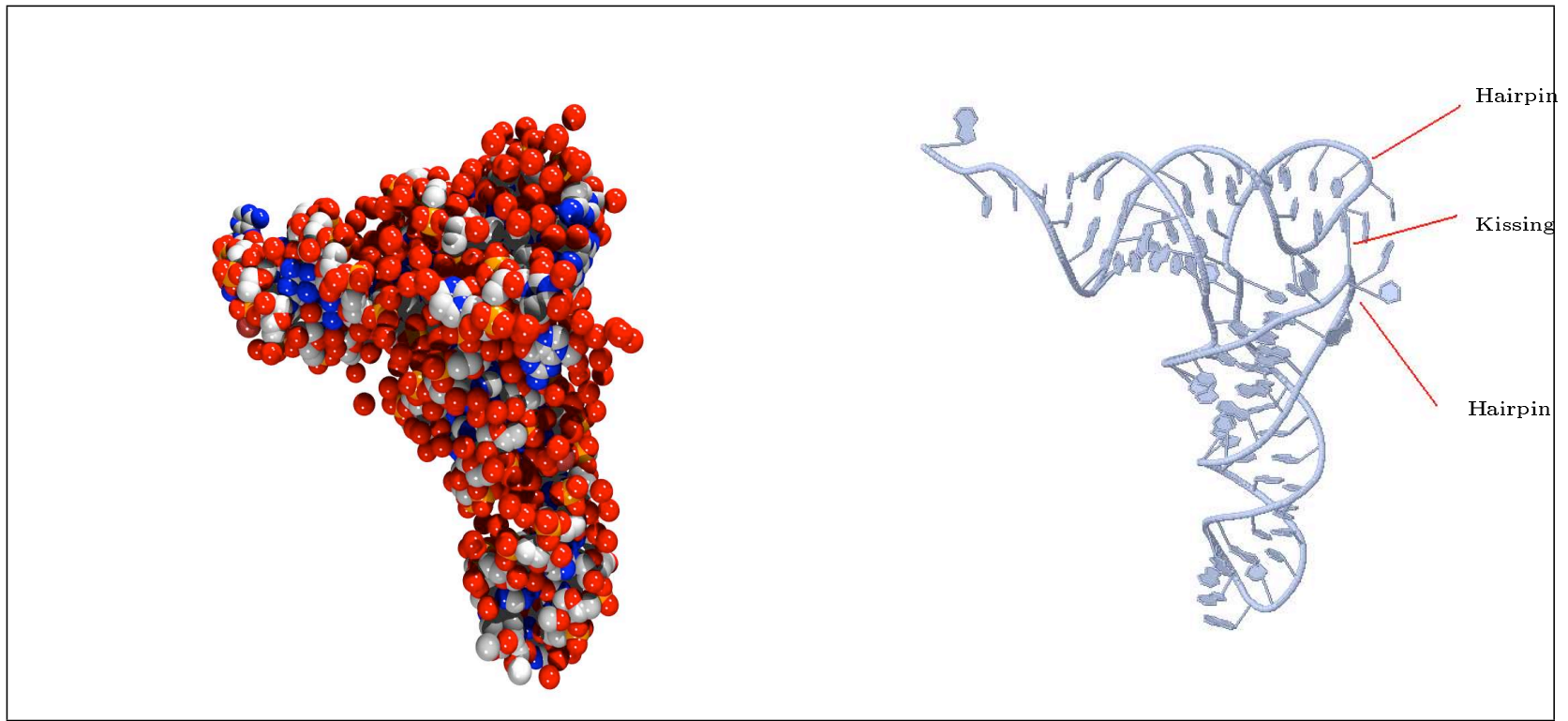
Figure 10: A typical tRNA (PDB ID 1evv [34]. It has the genus 1 of a kissing hairpin pseudoknot.
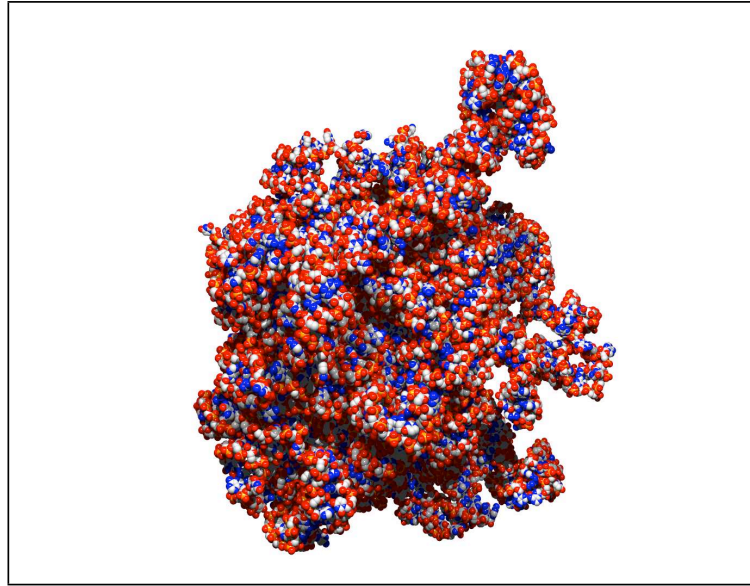
Figure 11: The B chain of 1vou.pdb is an RNA of genus 7 and of length 2825 bases.

- This PK of genus 7 is made of 3 HPK, 3 KHP nested in a large KHP

- Are these genii big?

Exact enumeration of RNA structures.

- Model: RNA in which any base can pair with any other base. All pairing energies are identical

$$V_{ij} = v$$

- Partition function of the model can be written as

$$Z_N(L) = \frac{1}{A} \int d\phi \; e^{-\frac{N}{2v} \mathrm{Tr}\phi^2} \frac{1}{N} \mathrm{Tr} \left(1 + \phi\right)^L$$

- with only one *NxN* matrix ϕ

- This integral can be calculated exactly using random matrix theory (orthogonal polynomials).

$$Z_N(L) = \sum_{g=0}^{\infty} \frac{a_L(g)}{N^{2g}}$$

- and the asymptotic behaviors are given by

$$a_L(g) \approx_{L \to \infty} K_g (1 + 2v)^L L^{3g - 3/2}$$

$$K_g = \frac{1}{3^{4g - 3/2} 2^{2g+1} g! \sqrt{\pi}}$$

- The total number of diagrams with any genus is given by

$$\mathcal{N} \approx_{L\to\infty} L^{L/2} \frac{e^{-L/2+\sqrt{L}-1/4}}{\sqrt{2}}$$

- the average genus is given by

$$< g >_L \approx 0.25L$$

- for real RNA, the largest genus we found is 18 for ribosomes (size around 3000 bp). The genus should be around 750.

- What about Steric Constraints?

Enumeration of self-avoiding RNA structures.

- Self-avoiding polymer on a cubic lattice
- Saturating attraction between nearest-neighbor monomers.
- Monte Carlo growth method allows to calculate accurately free energies.
- Length of chains up to 1200

$$< g > \approx 0.13 L$$

- Still much bigger than for real RNA: 390 for RNA of length 3000 instead of 18.
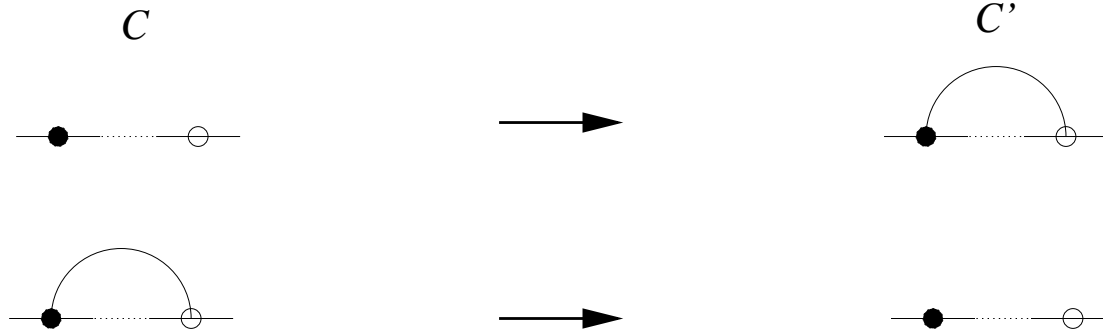
Monte Carlo method

- Idea: forget matrix fields, keep genus
- Work in pairing space (contact map)

$$Z = \sum_{\text{possible pairings}} e^{-\beta E(\text{pairing})}$$

- Introduce a chemical potential for the topology: $e^{-\mu} = \dfrac{1}{N^2}$

$$Z = \sum_{\text{possible pairings}} e^{-\beta E(\text{pairing}) - \mu g(\text{pairing})}$$
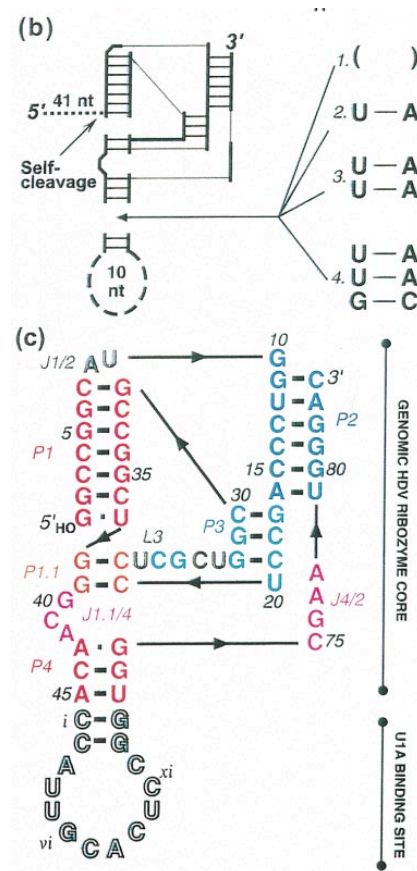
Possible moves

- Accept or reject move with probability

$$p = e^{-\beta \Delta E - \mu \Delta g}$$

- It is possible to
  - take into account the entropy
  - make it very fast
  - take into account steric constraint
- We are able to find the correct pseudo-knots in RNA up to size 200
  - transfer RNAs
  - Hepatitis delta virus ribozyme

The structure of the HDV ribozyme

Conclusion

- Matrix field theory introduces a natural classification of RNA folds according to their topological genus.

- One can write exact recursion equations for genus 0, 1, ...

- Most promising is the Monte Carlo calculation with chemical potential for the genus.