RNA folding
and
Matrix field Theory

Henri Orland (IPhT,CEA, Saclay)

Collaboration with

- A. Zee (KITP, UCSB)

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

Outline

- Review of basic properties of RNA
- Secondary structures
- Matrix field theory for RNA
- Topological classification of RNA
- Exact enumeration of RNA structures
- Monte Carlo approach

Review of basic properties of RNA

- RNA is a biopolymer
  - RNA (length ~ 70–3000): single stranded
  - DNA (length ~ 10^6–10^9): double stranded
  - Proteins (length ~ 10^2)
  - Polysaccharides (length ~ 10^3)

Several forms of RNA

- Messenger : mRNA (L ~ 1000) (only 5% of RNA)
- Transfer: tRNA (L ~ 70)
- Ribosomal: rRNA (L ~ 3000)
- Micro: μRNA (L ~ 25)
- Huge amounts of non-coding RNA in "junk" DNA: up to 80%

Why does the 3d structure of RNA matter?

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

Important discovery since 2000: some RNAs play crucial role in cell regulation

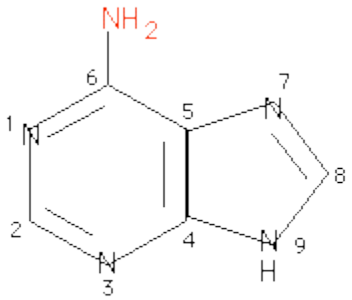
Function strongly related to shape

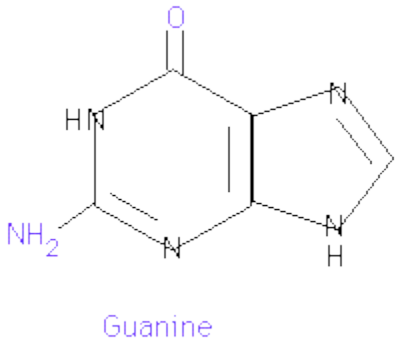
⟶ Must know 3d structure of RNA

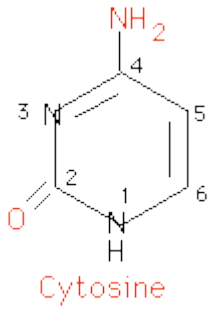
Chemistry of RNA

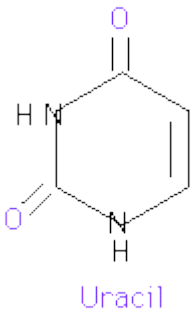
- RNA is a single-stranded heteropolymer
- Four bases:
  - Adenine (A)
  - Guanine (G)
  - Cytosine (C)
  - Uracil (U)

The sugar phosphate backbone polymerizes into a single stranded charged (-) polymer









Energy scales

- Crick-Watson: conjugate pairs

  C – G

  A – U

Pairing due to Hydrogen bonds between bases ⟹ RNA folding

Stacking of aromatic groups

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

Energy scales

C — G : 3kCal/mole = 5 kT

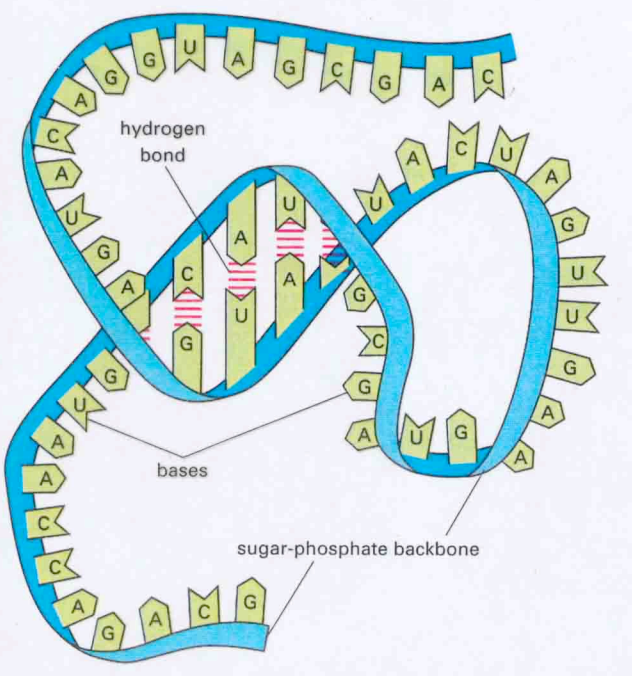
A — U : 2kCal/mole = 3.3 kT

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



RNA folding problem:
determine which
bases are paired

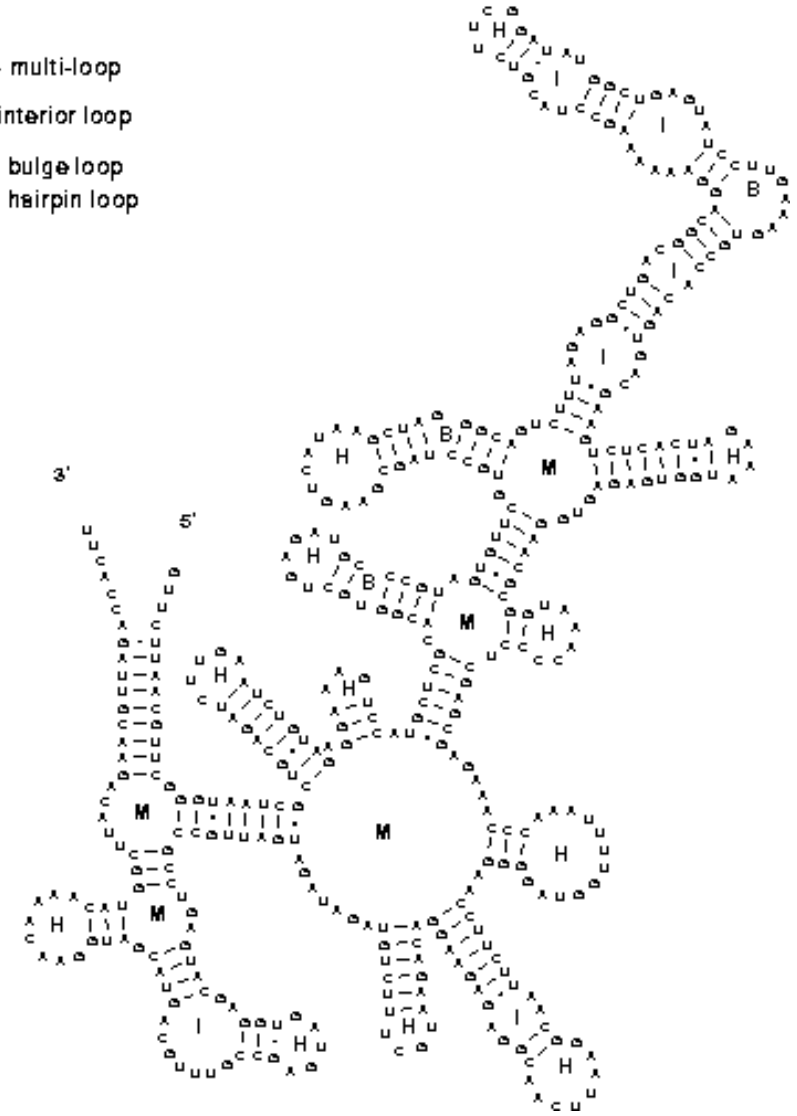
Bacillus subtilis RNase P RNA

M - multi-loop

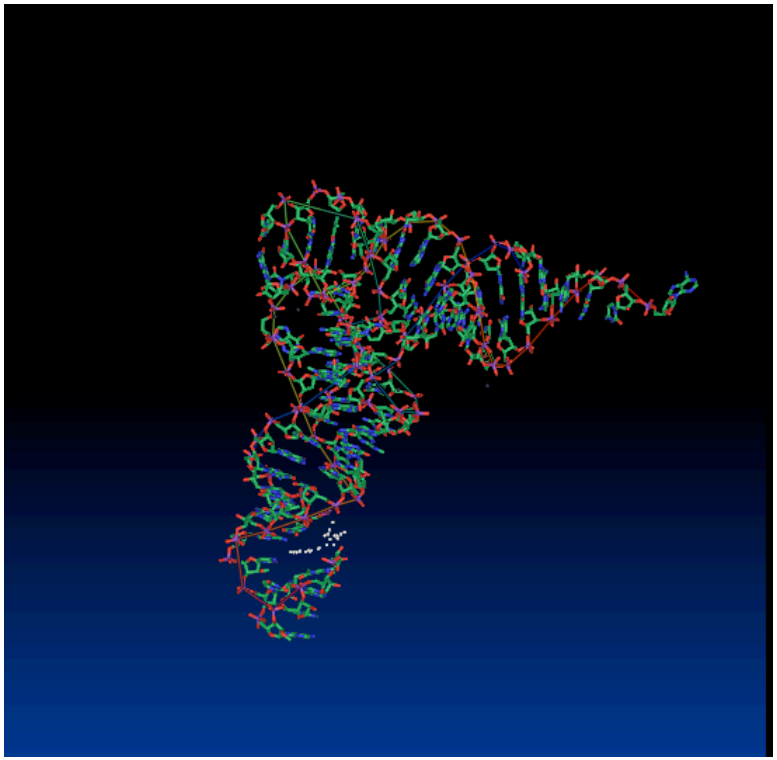
I - interior loop

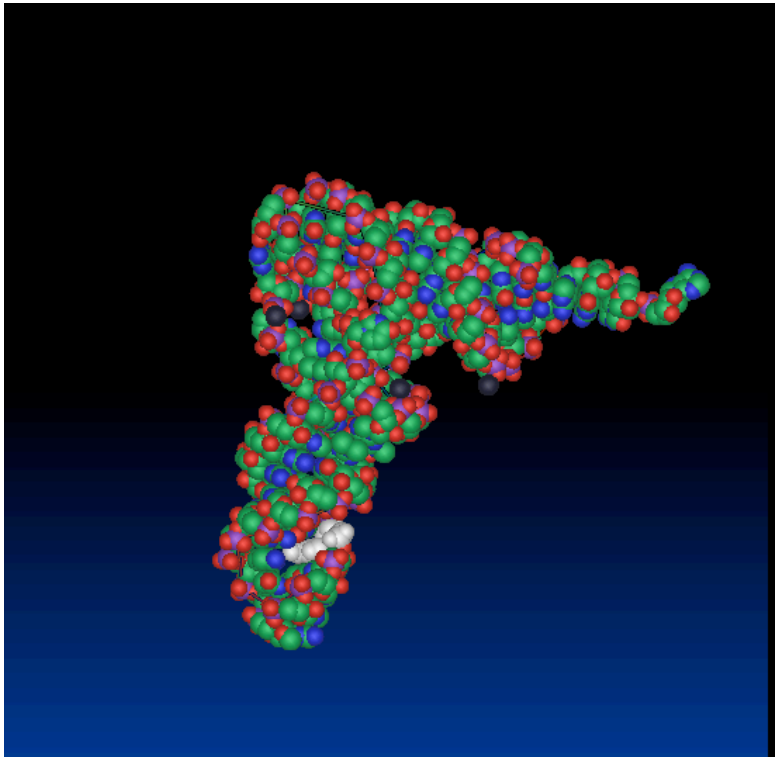
B - bulge loop

H - hairpin loop



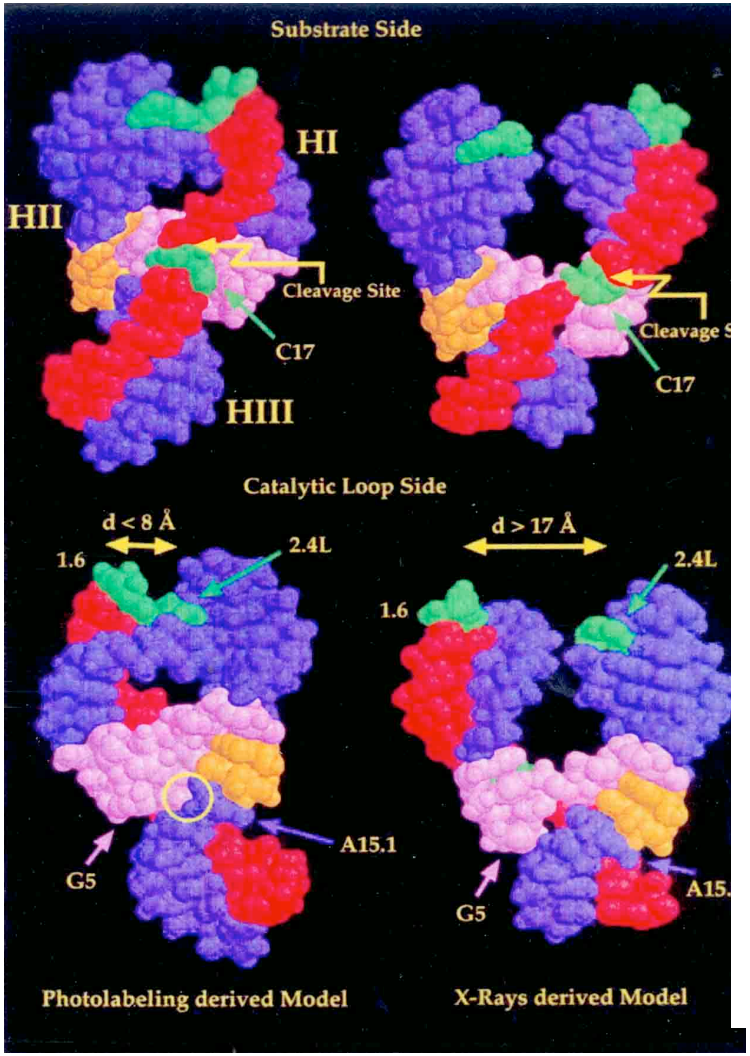
Pictures of RNA



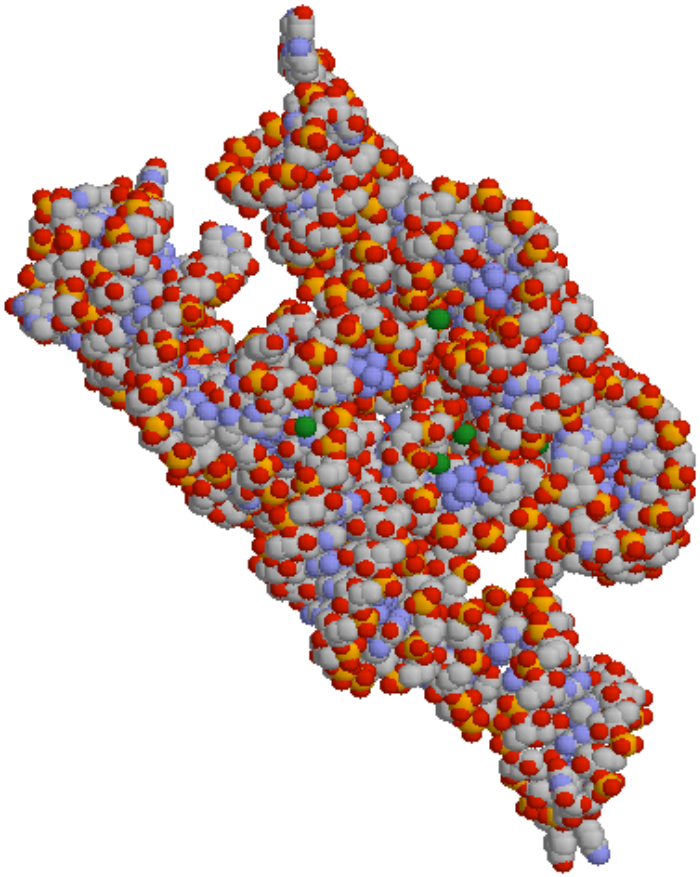


Transfer RNA

Hammerhead Ribozyme

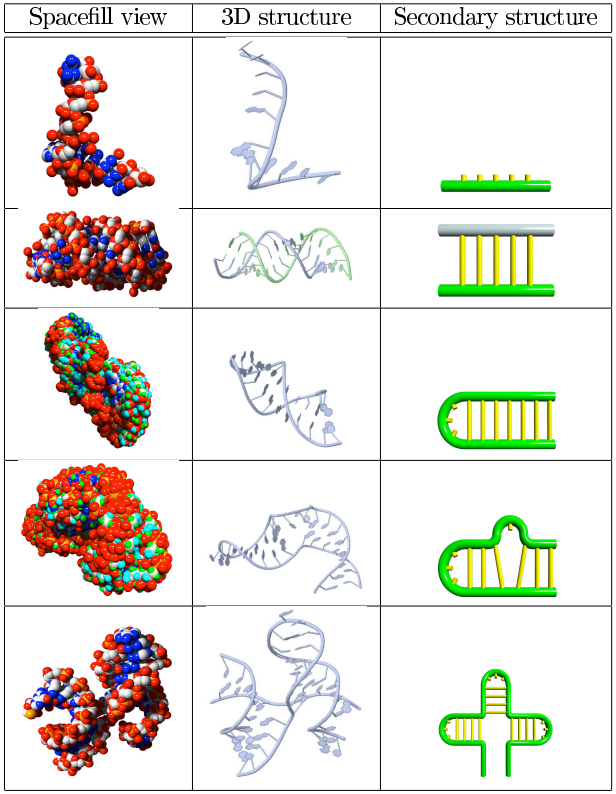


Ribosomal RNA



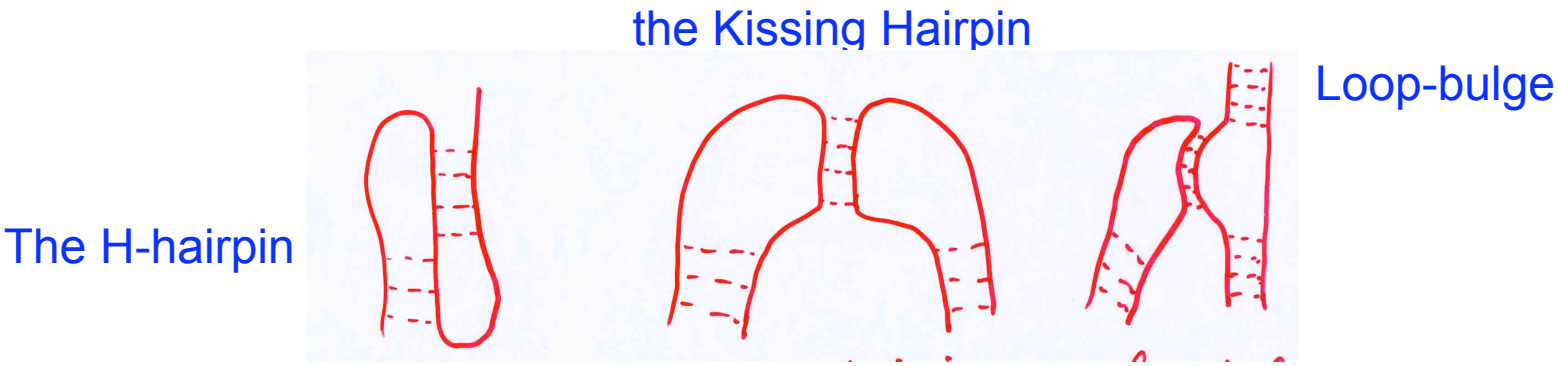
Secondary structures elements

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



Pseudo-knots in RNA

- In addition to planar structure,
- there are “pseudo-knots” which constrain the 3d structure



- 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
- experimental observation: number of bases in pseudo-knot << number of bases in planar secondary structure (less than 10%)

⇒ RNA folding much easier than protein folding

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

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

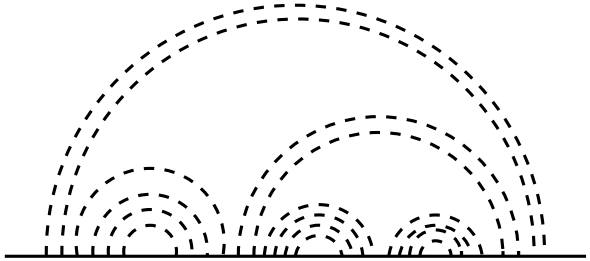
- In fact the $\log \mu$ goes into the free energies of pairing, so that

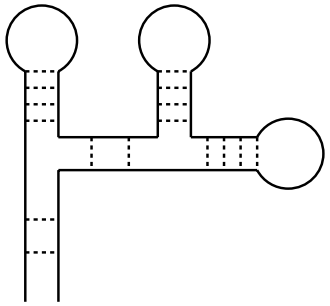
$$S = -c \log l$$

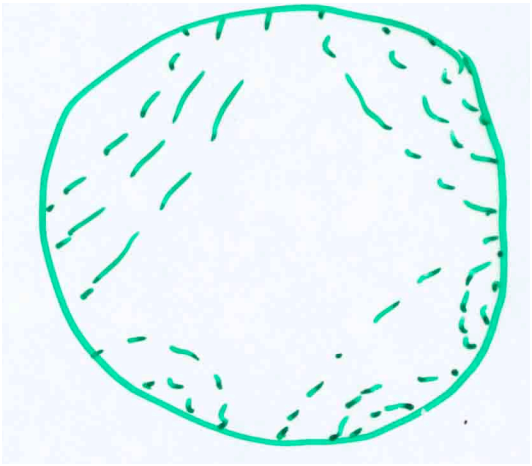
- with $c = 3/2$ (Gaussian chain)
- $c = 1.75$ (Self Avoiding Walk)

Planar Secondary structures

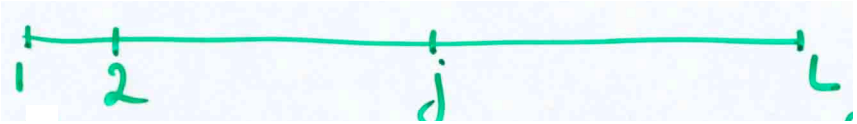
- We work on Q_0
- Secondary structures = Arches







- Define $Z(i,j)$ as the
- partition function of segment (i,j)



Recursion relation

- Graphically, when one adds a base

$$Z(i,k+1) = Z(i,k) + \sum_{j=i}^{k} V_{j,k+1} Z(i,j-1) Z(j+1,k)$$

- 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 : about 75% of base pairings correctly predicted
- Algorithm scales as N^3
- One can include Entropies and Stacking Energies
- MFOLD
- Vienna Package

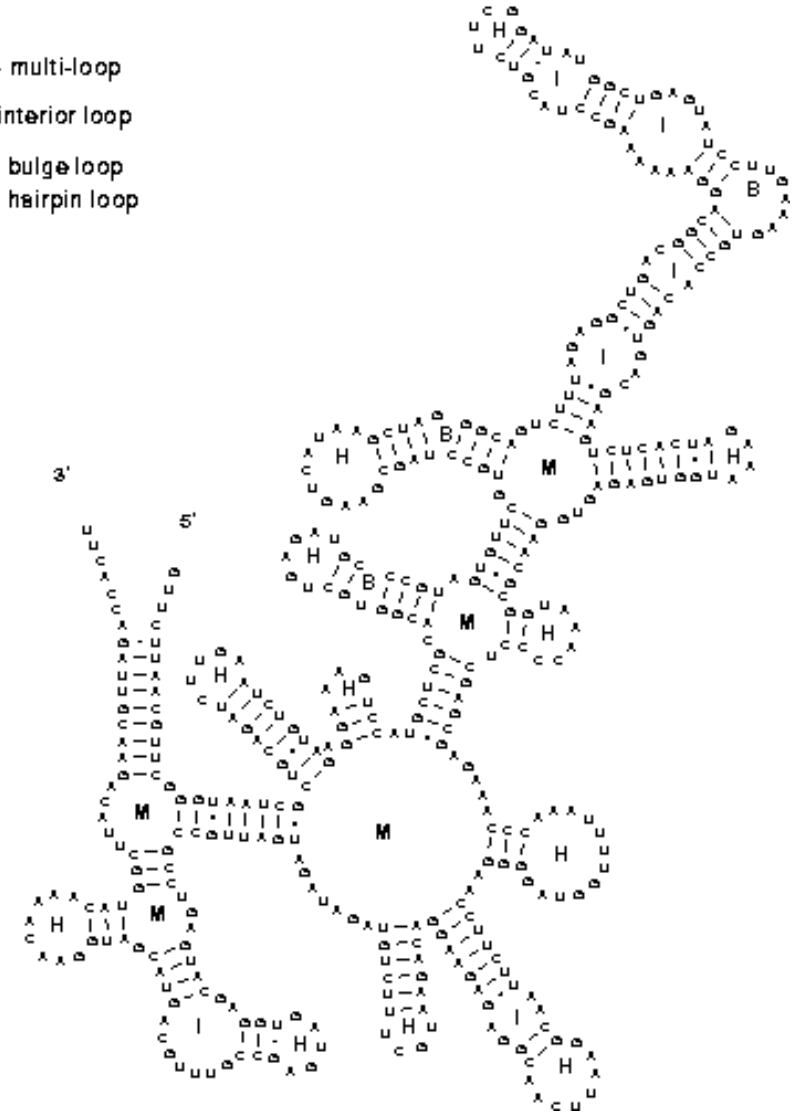
Bacillus subtilis RNase P RNA

M - multi-loop

I - interior loop

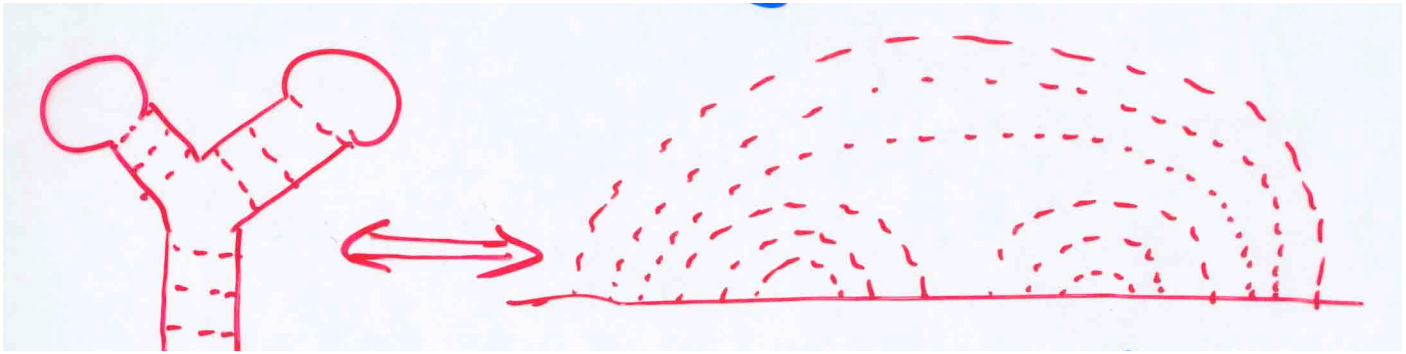
B - bulge loop

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

$$N \to +\infty \equiv$$ Secondary structures

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

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

$\frac{1}{N}$ plays the role of $[Mg^{++}]$

- 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.

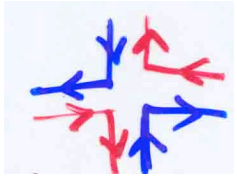
Matrix Field Theory

- 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: $NTr\phi^4_{ab}(x)$ ⟹



factor N

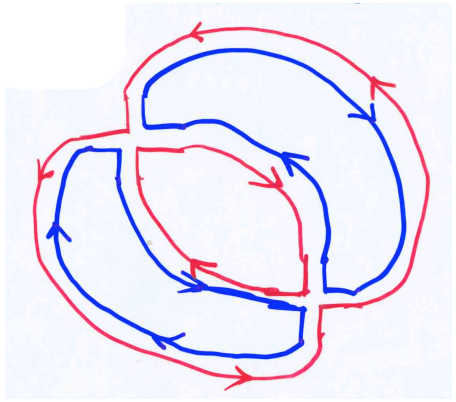
- Propagator: $\frac{1}{N}G(x-y)$ ⟹

$\frac{1}{N}$

Feynmann Graphs

- V: vertices
- I: internal propagators
- L: loops

$$N^{V-I+L}$$



- V=2
- I=4
- L=4
- Euler characteristic:

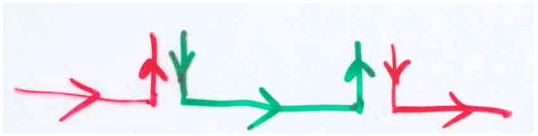
$$\chi = V - I + L$$

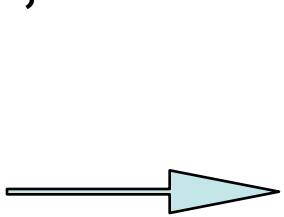
Euler characteristic and the Genus

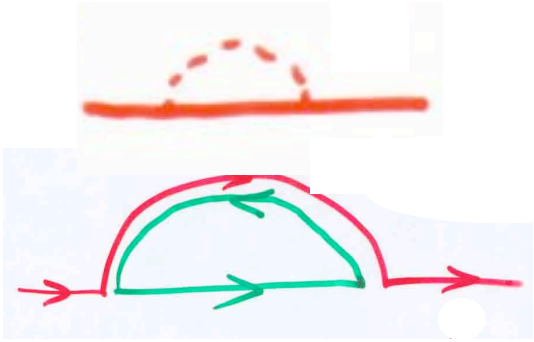
- Consider a graph with Euler characteristics χ
- Theorem: this graph can be drawn without crossings on a surface of genus given by $g = \frac{2 - \chi - c}{2}$ where c is the number of boundaries of the graph
- The genus g is the number of handles of the embedding surface

Double line graphs

- In our problem, if we use matrix fields





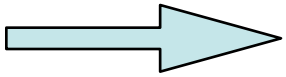


$\phi_{ab}(x)$: NxN matrix

- If we use same rule:

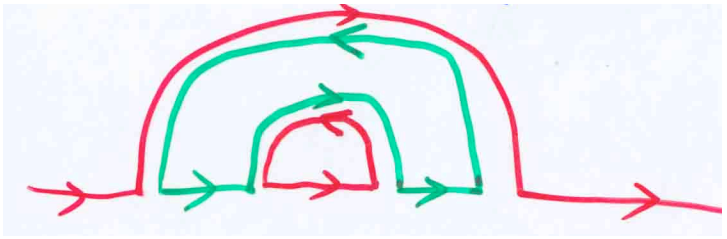
Propagator: 1/N
Loop: N

- Above graph:



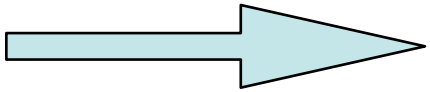
$$N \times \frac{1}{N} = 1$$

- Other graph



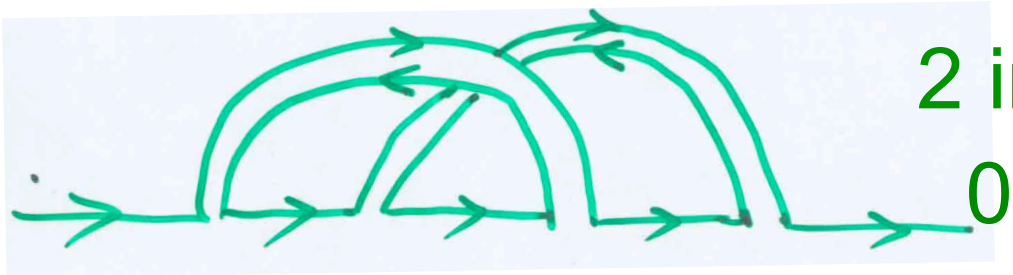
2 internal lines: $1/N^2$

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

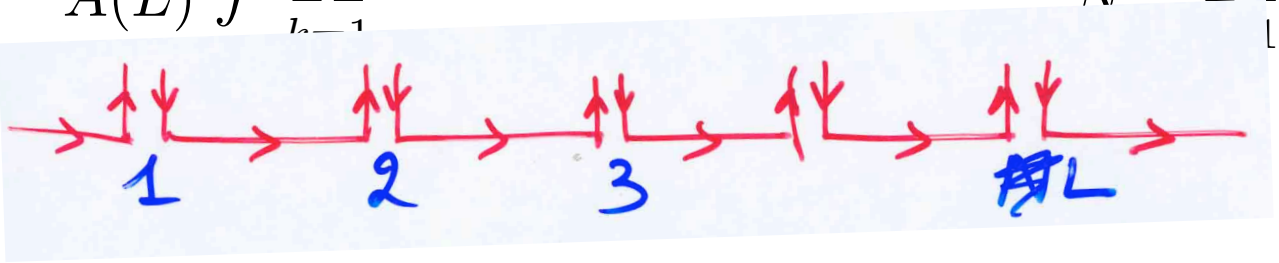
- Matrix field theory seems to do what we want.
- Not really surprising:

Hollywood already knew: the MATRIX rules

Matrix field representation of RNA folding

- We thus generalize the Wick theorem

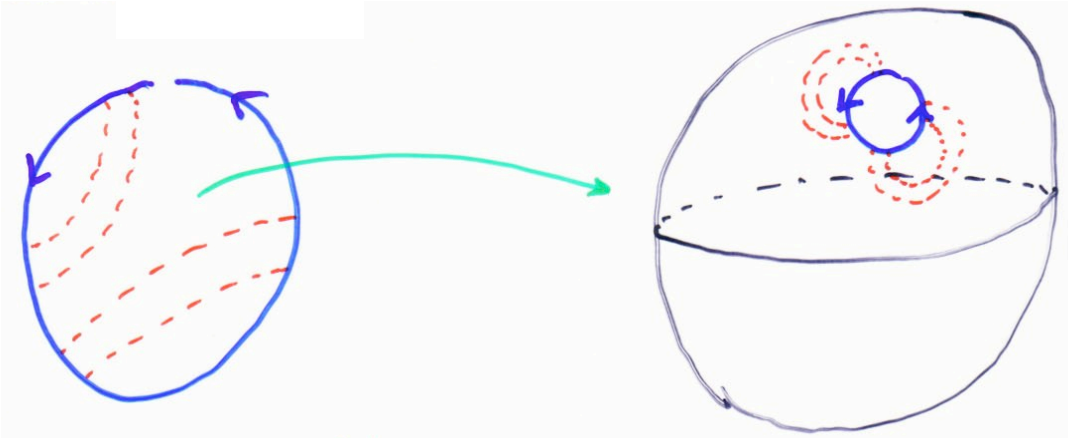
$$Z(1,L) = \frac{1}{A(L)} \int \prod_{k=1}^{L} d\varphi_k e^{-\frac{N}{2}\sum_{ij}(V^{-1})_{ij}\,\mathrm{tr}(\varphi_i\varphi_j)} \frac{1}{N}\,\mathrm{tr}\prod^{L}(1+\varphi_l)$$



- $\varphi_{ab}(i)$ is a real symmetric (or hermitian) matrix
- By looking at a few diagrams: 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:



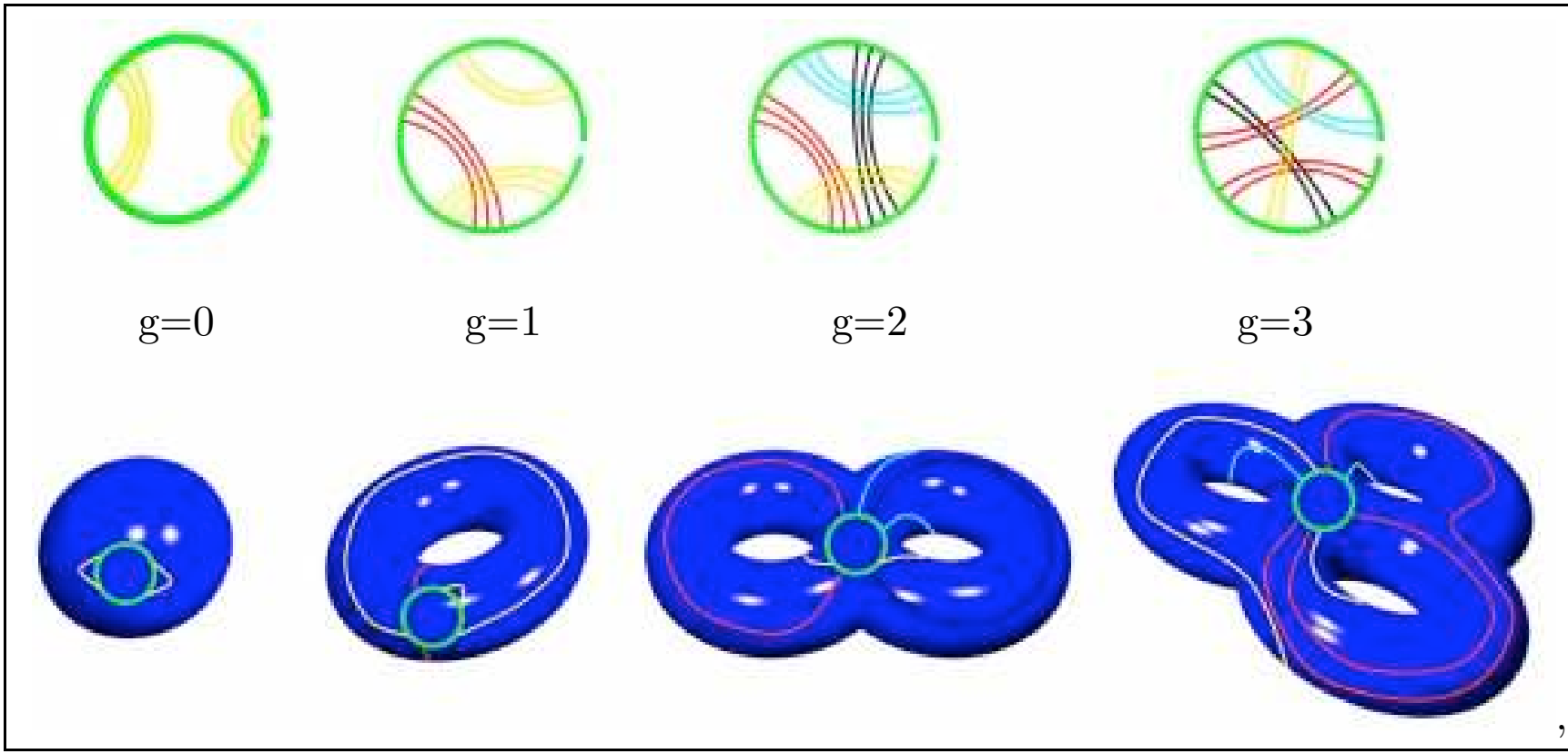
- Number of handles of embedding surface

$$g = \frac{P - L}{2}$$

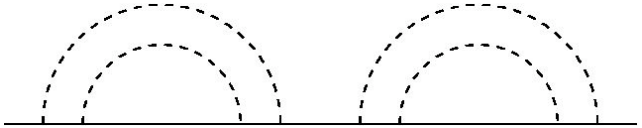
- In fact, one can prove that the matrix field partition function is equal to

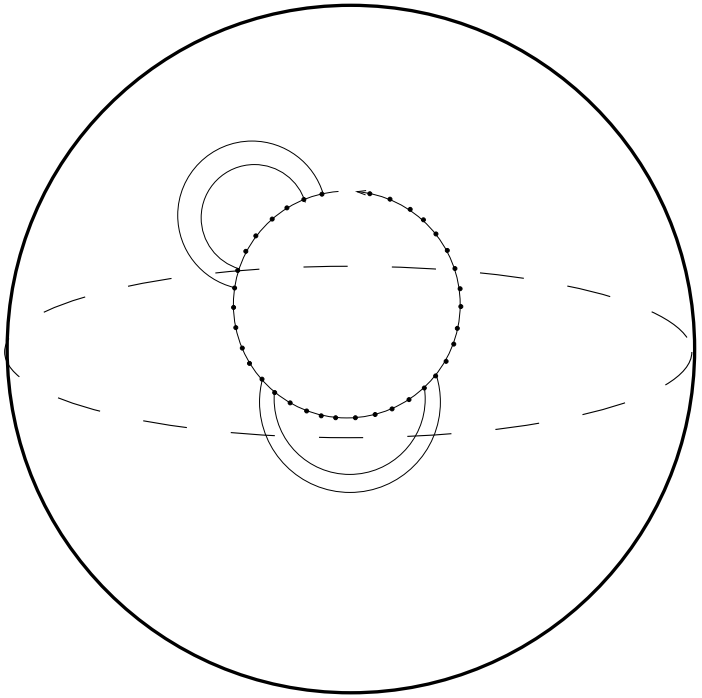
$$Z = \sum_{\text{all pairings}} \frac{1}{N^{2g(\text{pairing})}} e^{-\beta E(\text{pairing})}$$

- where *g*(pairing) is the genus of the pairing graph

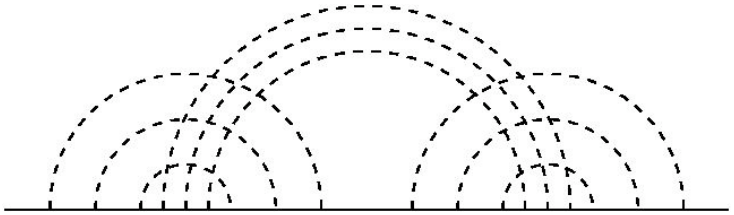

g=0
g=1
g=2
g=3
,,,

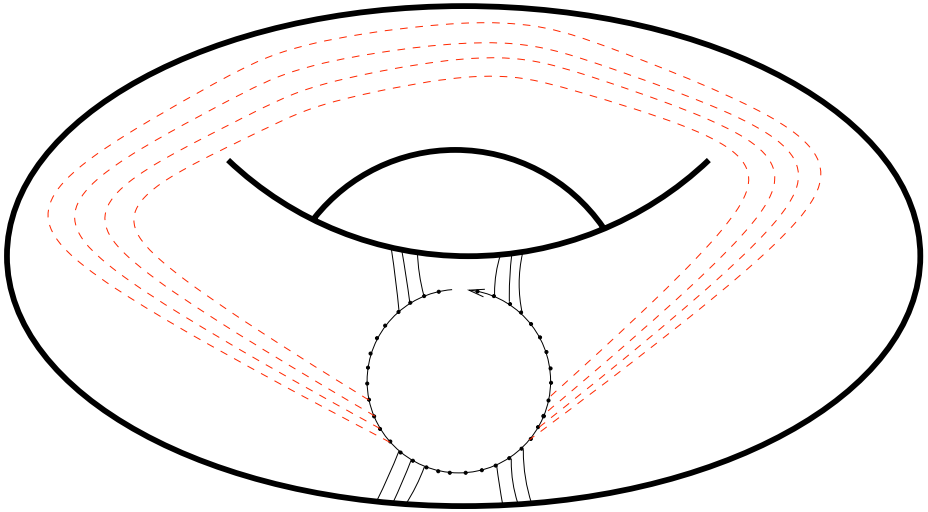
Genus 0: the Sphere



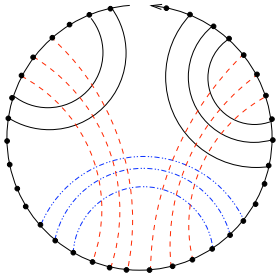


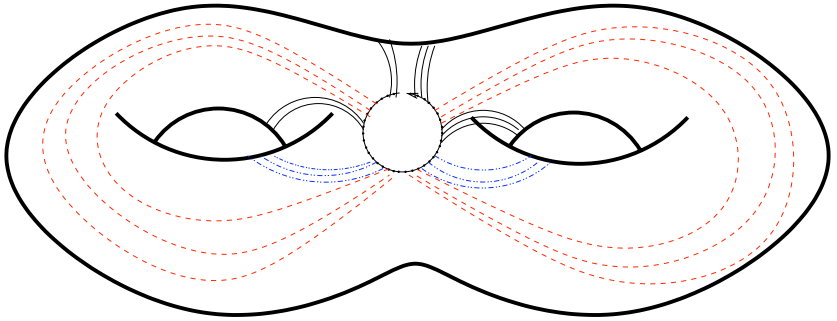
Genus 1: the Torus

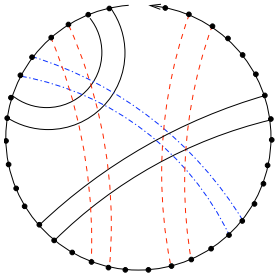


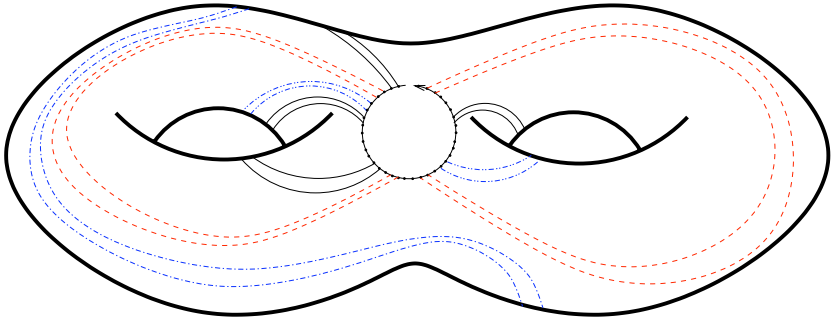


Genus 2: the Bi-torus

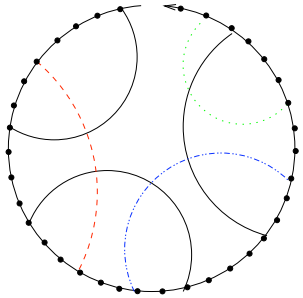


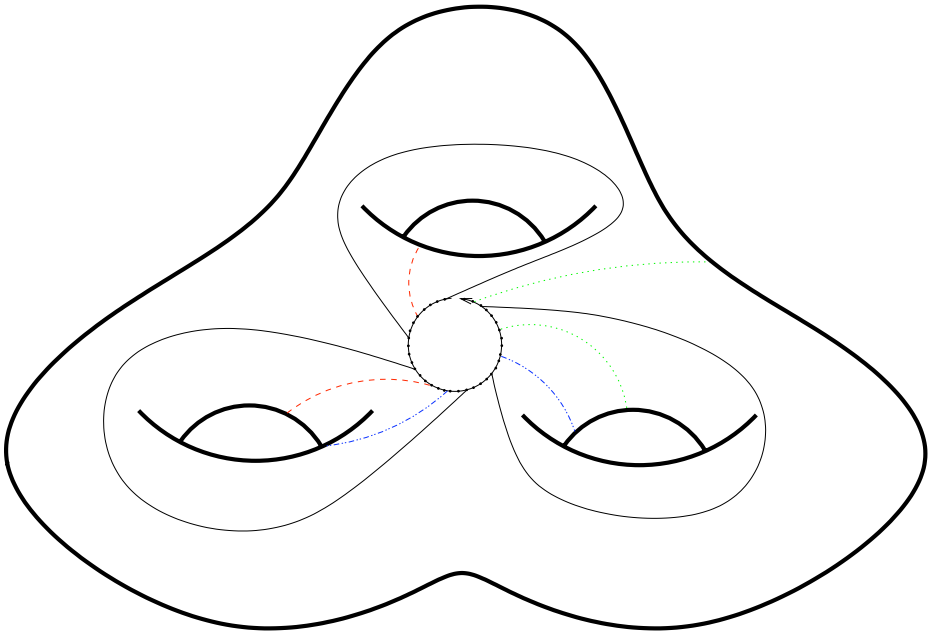






Genus 3





Large N expansion

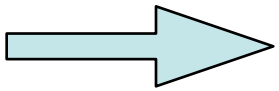
- After some algebraic manipulations, one has the exact expression:

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

- 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



one can perform a loop expansion (saddle-point)

The loop expansion

- Saddle-point equation

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

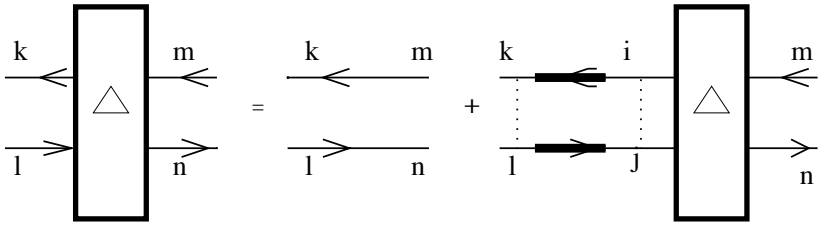
- Expansion in 1/N

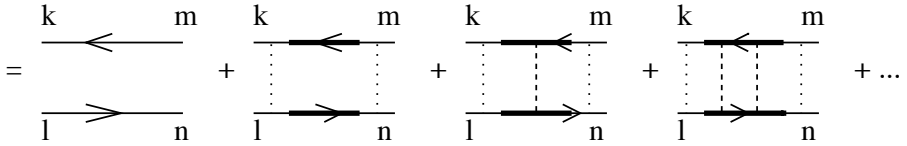
$$A_{ll'} = A^{(0)}_{ll'} + \frac{x_{ll'}}{\sqrt{N}}$$

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

Bethe-Salpeter equation

- No order 1/N correction





Recursion relations

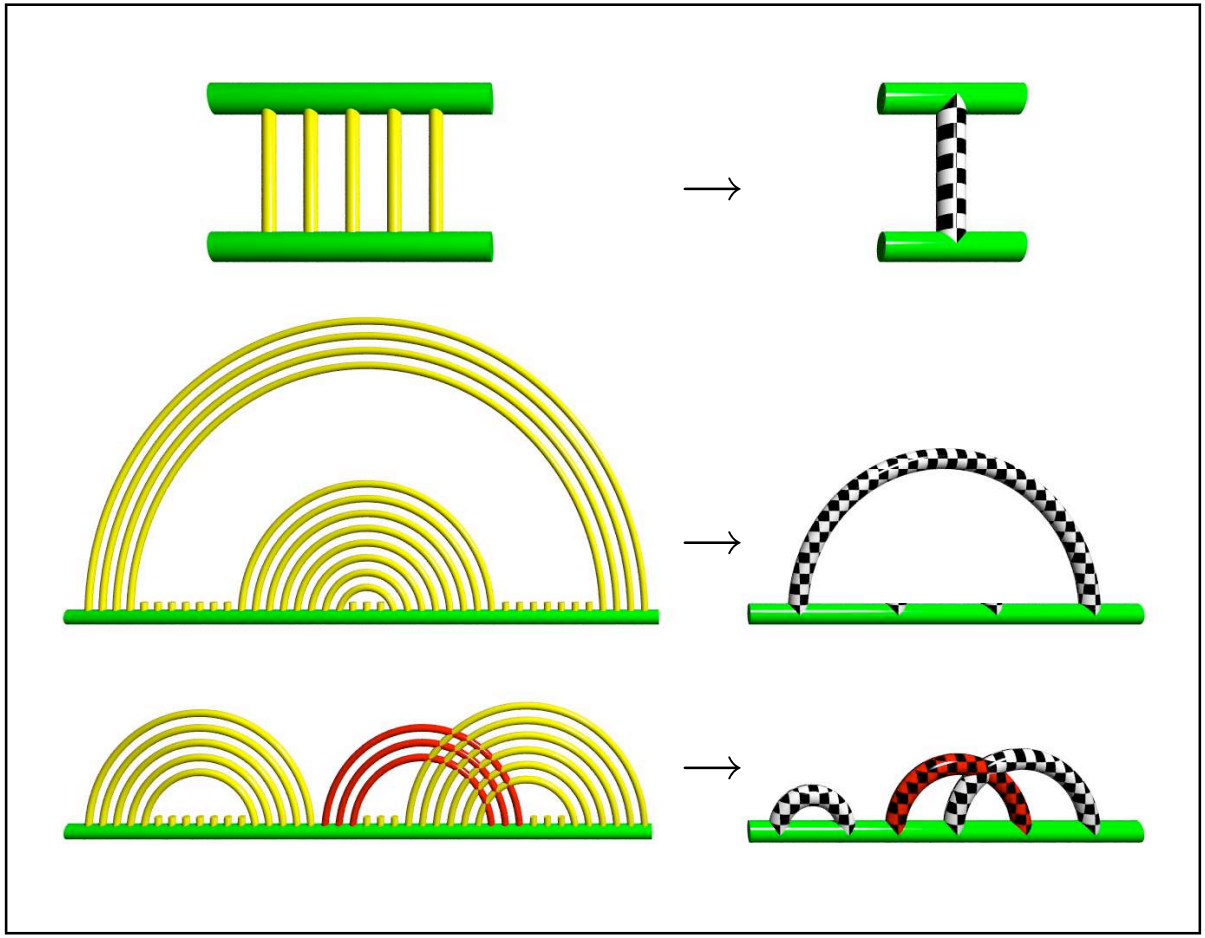
- It is possible to obtain exact recursion relations for genus 1
- There is an exact relation

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

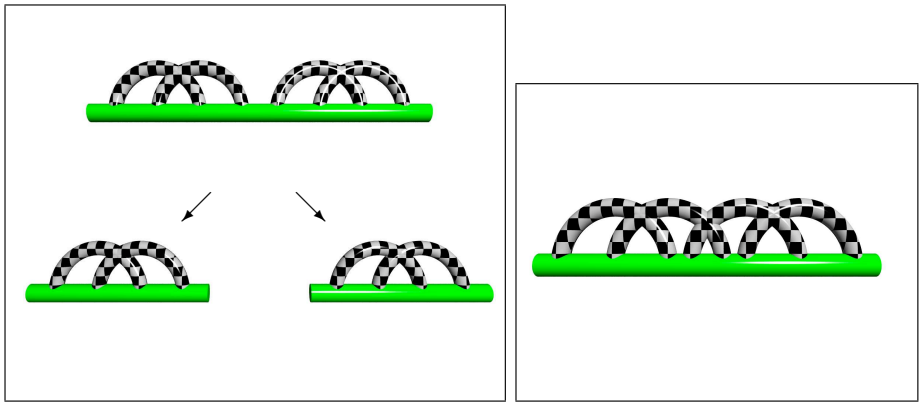
- which can be expanded in powers of $\frac{1}{N}$
- Algorithm scales as L^6 ⟶ too long!

Graphology

Parallel pairings don't change the genus

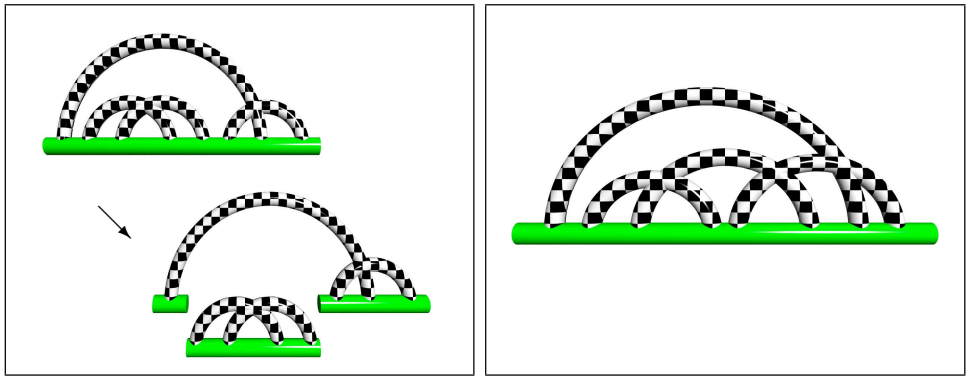


Irreducibility and Nesting



Irreducible PK

Genus is additive

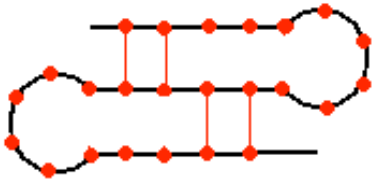


Non nested PK

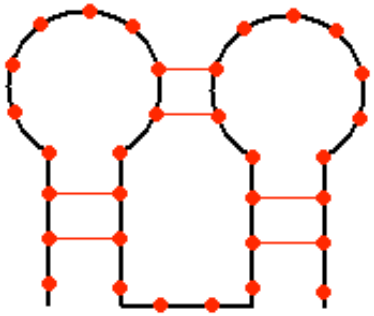
Only 4 primitive PK of genus 1

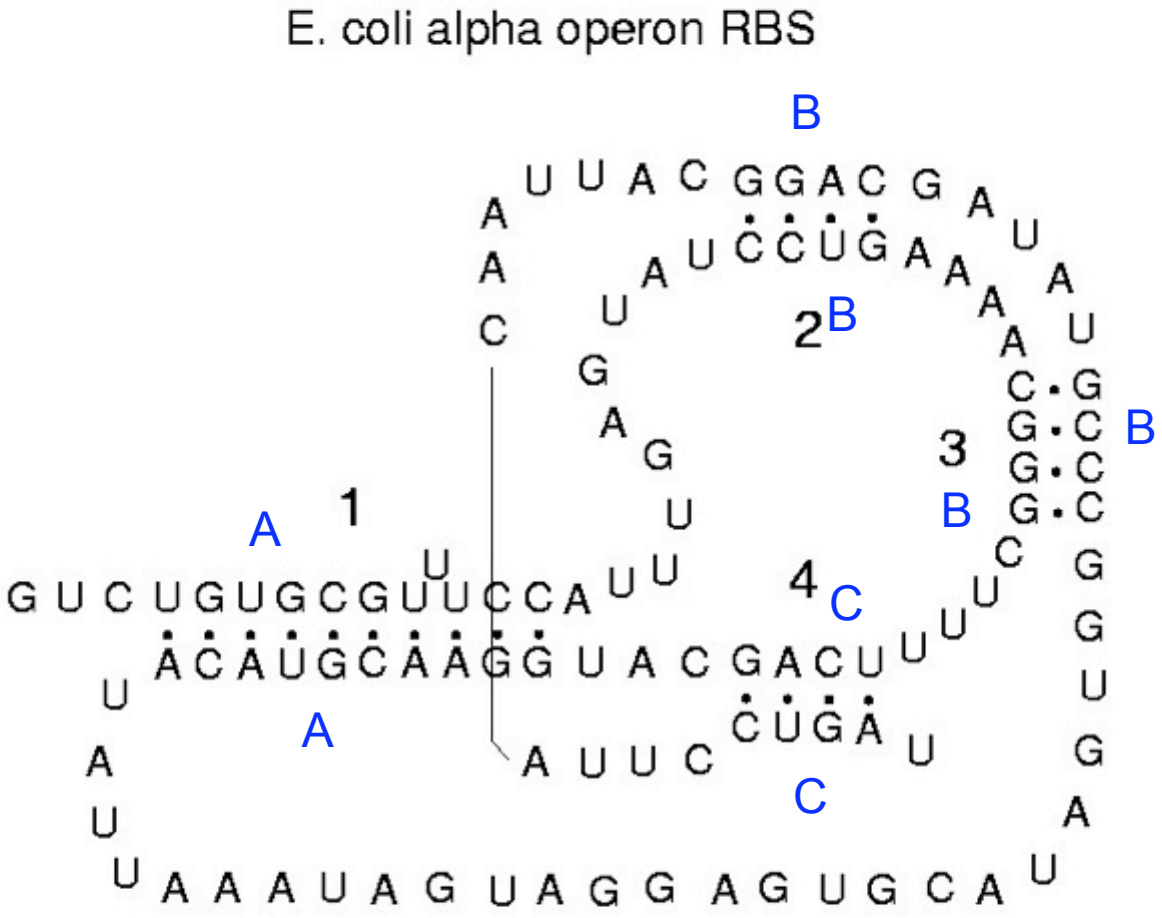
Primitive=Irreducible and non-nested

H PK



KHP




E. coli alpha operon RBS
B
2B
3
B
1
A
4
C
A
C
ABCABC

EMBL Accession number	Description	Start	End	Bits Score	Get Seq
AB032408.1	Shewanella violacea genes for SecY, ribosomal protein S13, ribosomal ...	554	653	69.9400	☐
AE004325.1	Vibrio cholerae O1 biovar eltor str. N16961 chromosome I, section 233...	10782	10672	80.2700	☐
AE005556.1	Escherichia coli O157:H7 EDL933 genome, contig 3 of 3, section 175 of...	8956	8845	102.2000	☐
AE006177.1	Pasteurella multocida subsp. multocida str. Pm70 section 144 of 204 o...	3108	2993	88.4700	☐
AE008857.1	Salmonella typhimurium LT2, section 161 of 220 of the complete genome...	15653	15542	100.2500	☐
AE014003.1	Yersinia pestis KIM section 403 of 415 of the complete genome.	5850	5961	97.1300	☐
AE015343.1	Shigella flexneri 2a str. 301 section 306 of 412 of the complete geno...	6155	6044	102.2000	☐
AE015474.1	Shewanella oneidensis MR-1 section 23 of 457 of the complete genome.	23	120	68.2200	☐
AE016767.1	Escherichia coli CFT073 section 13 of 18 of the complete genome.	243541	243430	102.2000	☐
AE016799.1	Vibrio vulnificus CMCP6 chromosome I section 3 of 11 of the complete ...	152117	152009	77.1800	☐
AE016848.1	Salmonella enterica subsp. enterica serovar Typhi Ty2, section 15 of ...	21171	21282	100.2500	☐
AE016992.1	Shigella flexneri 2a str. 2457T section 15 of 16 of the complete geno...	262117	262228	102.2000	☐
AE017127.1	Yersinia pestis biovar Medievalis str. 91001 section 1 of 16 of the c...	231109	231220	97.1300	☐
AE017156.1	Haemophilus ducreyi strain 35000HP section 6 of 6 of the complete gen...	132595	132483	79.8900	☐
AJ414141.1	Yersinia pestis strain CO92 complete genome; segment 1/20	232397	232508	97.1300	☐
AL627282.1	Salmonella enterica serovar Typhi (Salmonella typhi) strain CT18, com...	5691	5802	100.2500	☐
AP002564.1	Escherichia coli O157:H7 DNA, complete genome, section 15/20.	266263	266152	102.2000	☐
AP005073.1	Vibrio parahaemolyticus DNA, chromosome 1, complete sequence, 1/11.	278188	278299	72.3400	☐
AP005331.1	Vibrio vulnificus YJ016 DNA, chromosome I, complete genome, section 2...	149281	149389	77.1800	☐
BX571874.1	Photorhabdus luminescens subsp. laumondii TTO1 complete genome; segme...	272732	272618	82.0900	☐
BX950851.1	Erwinia carotovora subsp. atroseptica SCRI1043, complete genome	4490616	4490505	86.6200	☐
CR378663.1	Photobacterium profundum SS9; segment 1/12	343376	343482	61.9700	☐
X02543.1	E. coli alpha ribosomal protein operon for ribosomal proteins S13, S1...	141	252	102.2000	☐
M12432.1	E.coli alpha operon ribosomal protein S13 (rpsM) gene, 5' end and pro...	572	683	102.2000	☐
U18997.1	Escherichia coli K-12 chromosomal region from 67.4 to 76.0 minutes.	223298	223187	102.2000	☐
U32762.1	Haemophilus influenzae Rd KW20 section 77 of 163 of the complete geno...	6695	6810	91.4500	☐
U00096.2	Escherichia coli K-12 MG1655 complete genome.	3440572	3440461	102.2000	☐

27 potential ABCABC

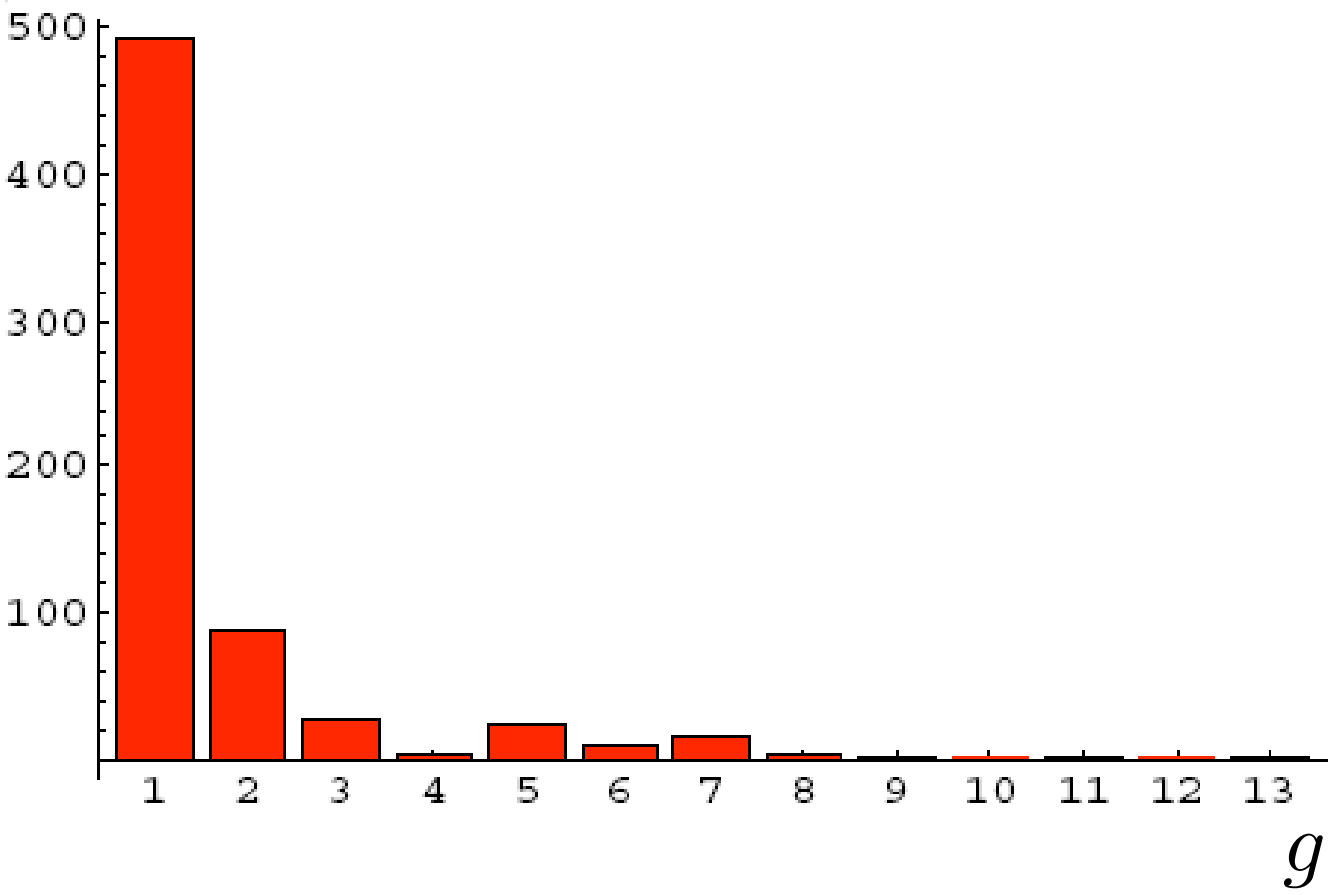
H. Orland, SPhT, Saclay

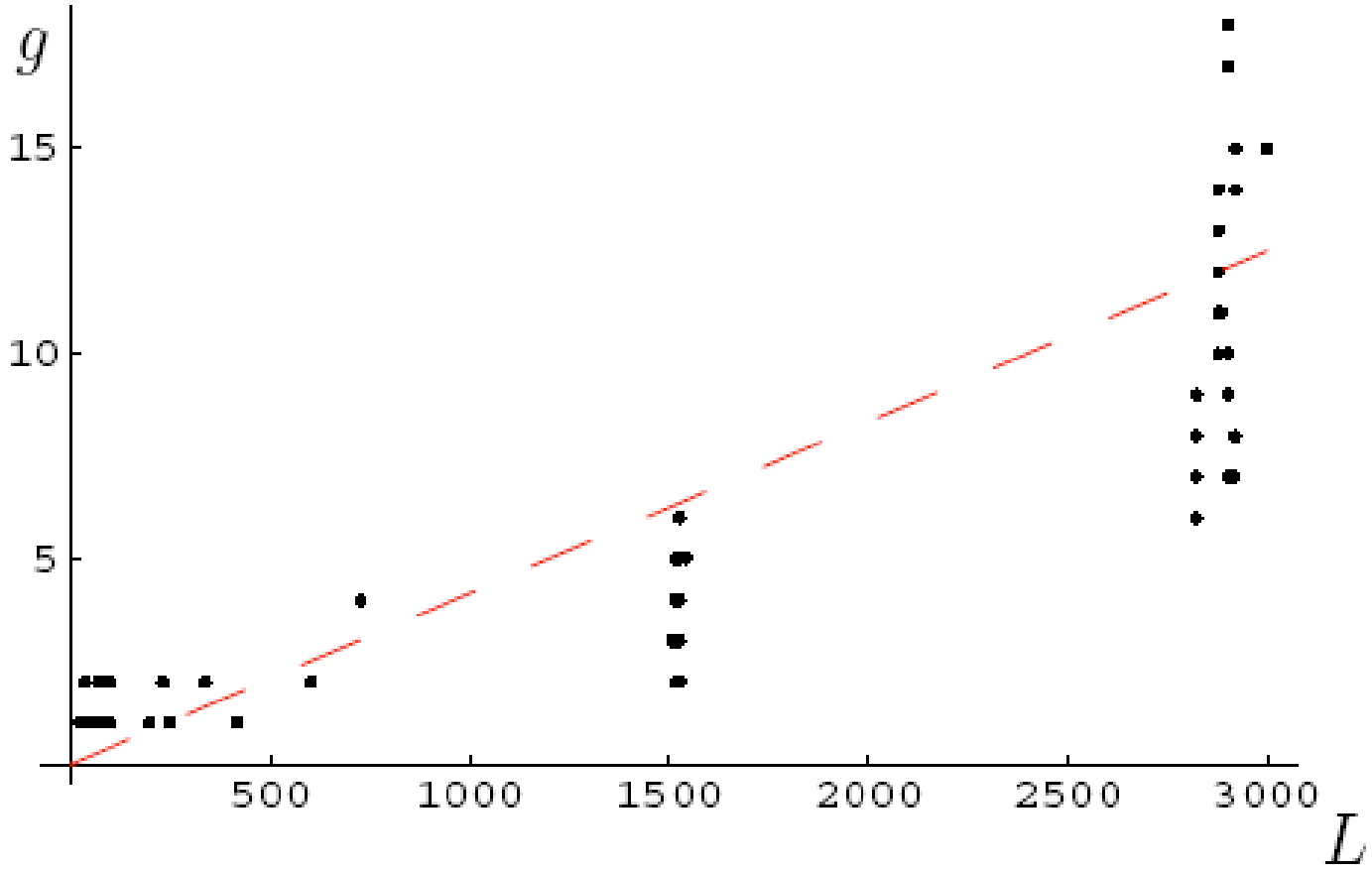
Statistical study

- Look in database and calculate genii of pseudo-knots
- PseudoBase: around 245 primitive pseudo-knots found experimentally
- 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
- 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

Number of RNA as a function of genus




g
15
10
5
500
1000
1500
2000
2500
3000
L

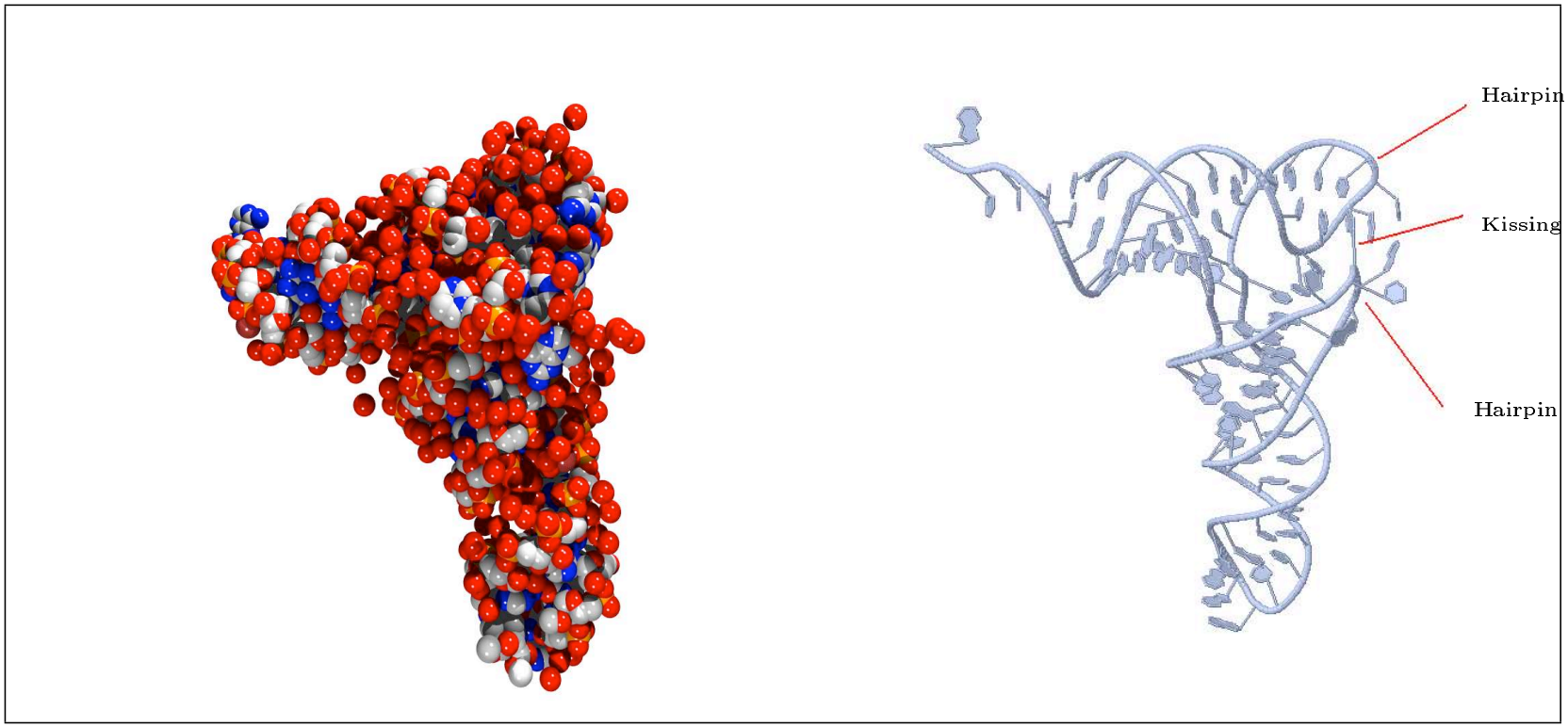


Figure 10: A typical tRNA (PDB ID 1evv [34]. It has the genus 1 of a kissing hairpin pseudoknot.

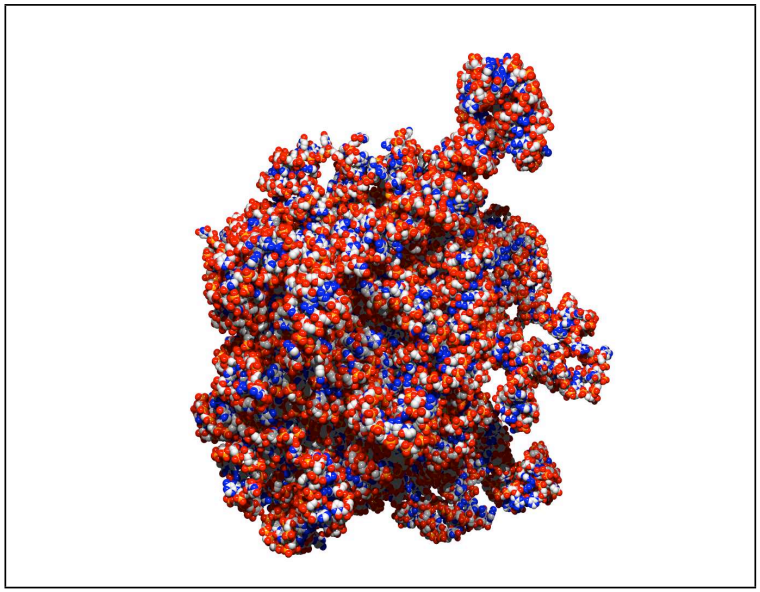
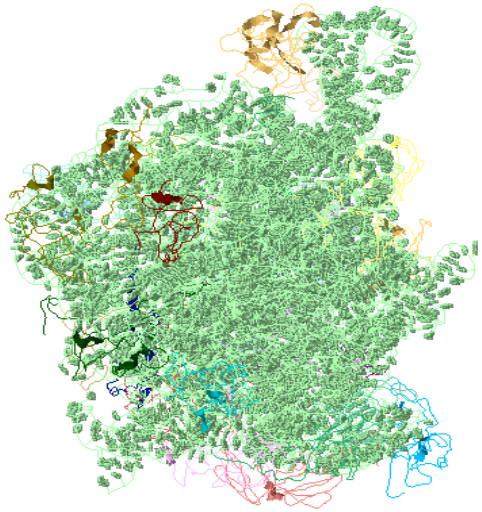


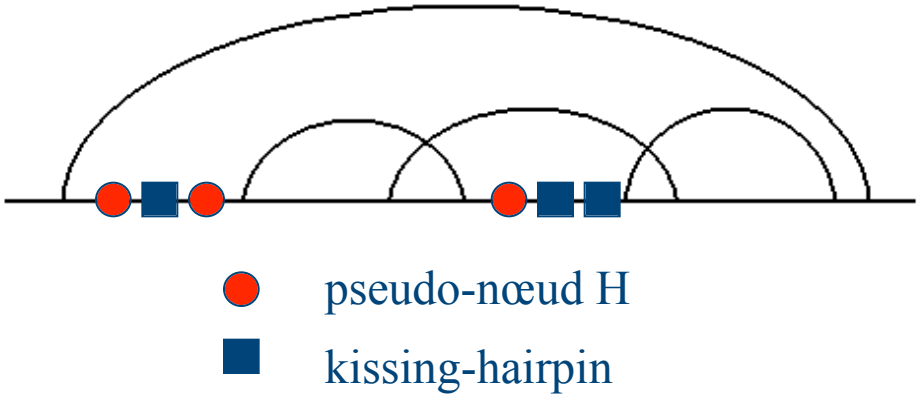
Figure 11: The B chain of 1vou.pdb is an RNA of genus 7 and of length 2825 bases.

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

Exemple de 1vou (sous-unité 50s d'un ribosome 70s d' E.Coli)



2850 bases



- This PK of genus 7 is made of 3 HP, 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

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

- with only one NxN matrix ϕ

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

number of graphs of length L of genus g

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

- 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

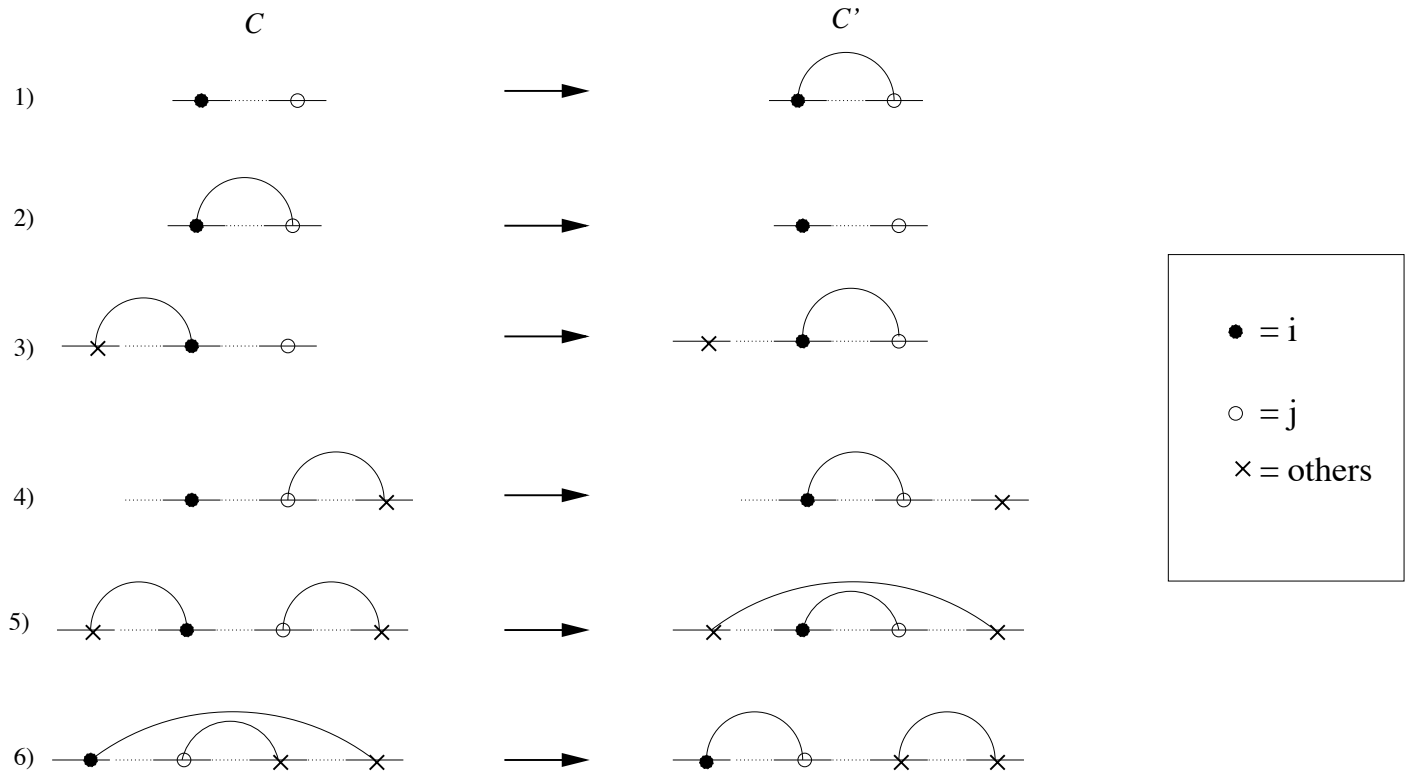
- Idea: forget matrix fields, keep genus
- Work in pairing space (contact map)

$$Z = \sum_{\text{possible pairings}} e^{-\beta E(\text{pairing})} / N^{2g(\text{pairing})}$$

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

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

Possible moves



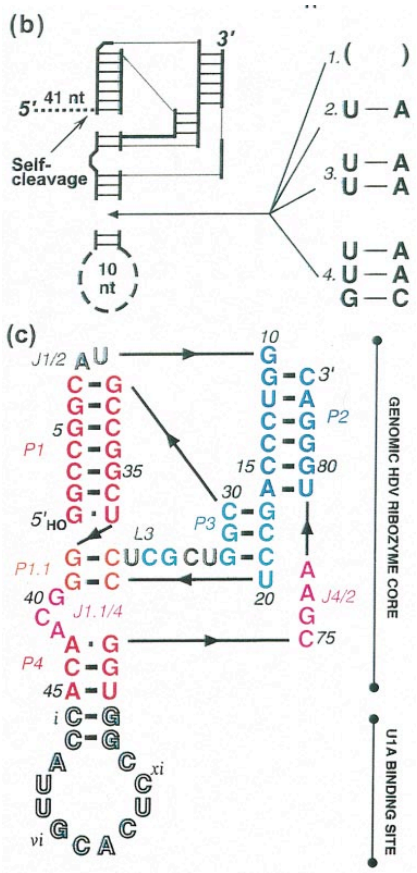
When a pair is added or removed, the energy is changed and the genus of the graph may have changed

- 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
- Current Turner energies are not fitted to calculate energies of pseudoknots
  - transfer RNAs (g=1)
  - Hepatitis delta virus ribozyme (g=2)

The structure of the HDV ribozyme



- We find pseudoknots with free-energies lower than the native ones!
- We have worked out a new parametrization of pairing free energies in RNA.
  - we predict 99% of all tRNA
  - predict very well RNA smaller than 200.

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, ...
- The Genus of real RNA is small
- Most promising is the Monte Carlo calculation with chemical potential for the genus (in fact enumerations of structures)