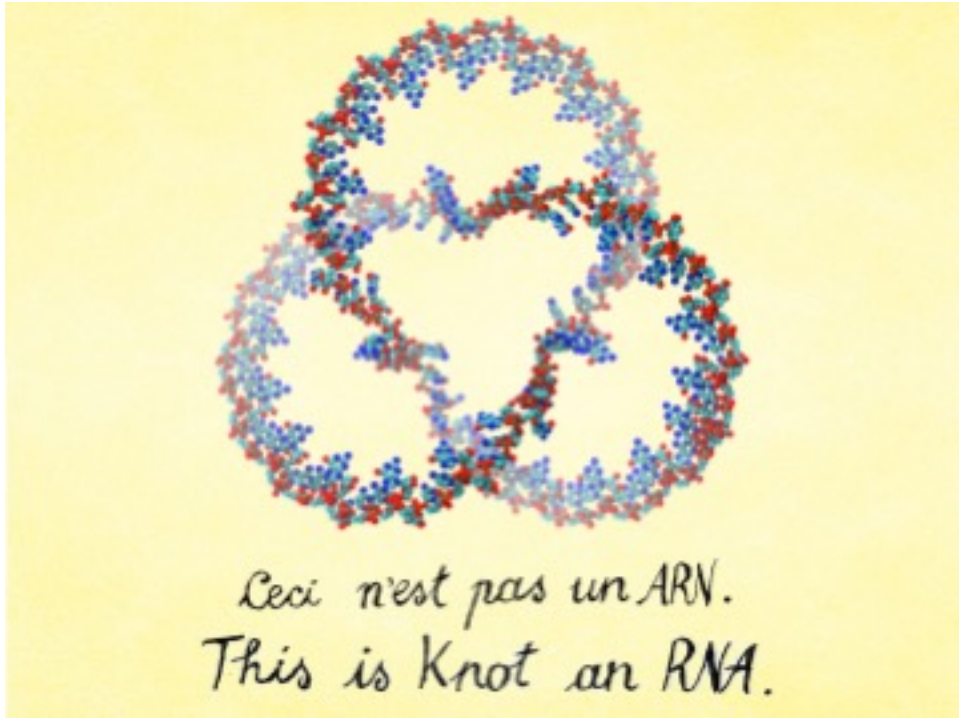
Pseudoknots and Knots in RNA



Henri Orland
Institut de Physique Théorique,
CEA-Saclay
France

A collaboration with

A. Zee (UCSB)

G. Vernizzi (Sienna College)

M. Bon (IPhT, Saclay)

C. Micheletti (SISSA, Italy)

Web server: http://ipht.cea.fr/rna/mcgenus.php

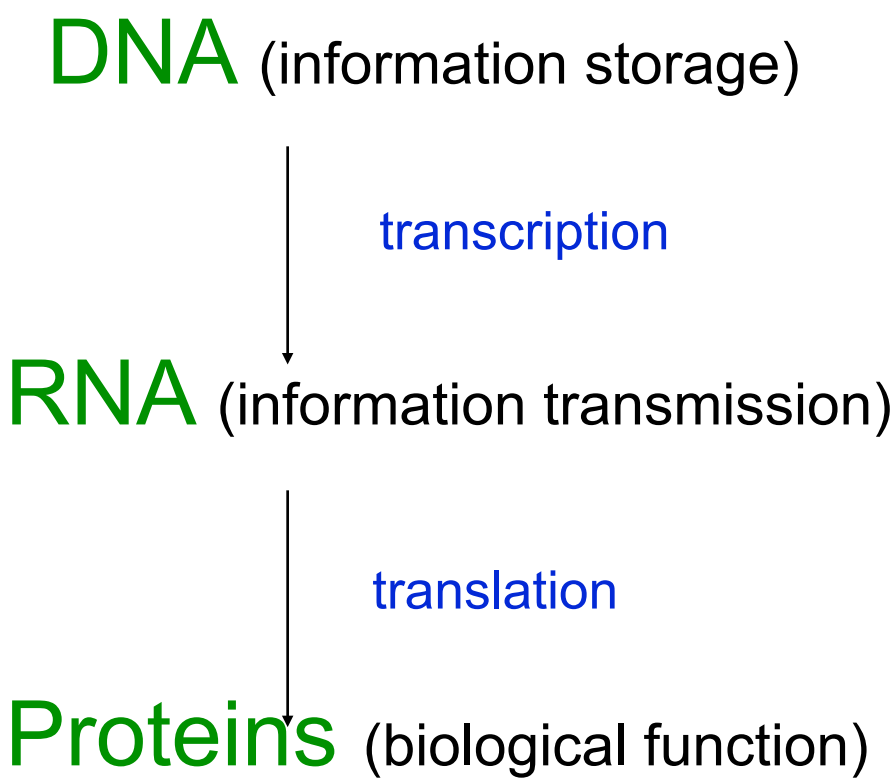
Outline

- Some basic properties of RNA
- Secondary structures
- Matrix field theory for RNA
- Topological classification of RNA
- Exact enumeration of RNA structures
- Algorithms for prediction
- What about knots?

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)

Central dogma of Biology



Several forms of RNA

- Messenger : mRNA (L ~ 1000) (only 5% of RNA)
- Transfer: tRNA (L ~ 70)
- Ribosomal: rRNA (L ~ 3000)
- Micro: μRNA (L ~ 25)
- Small interfering RNA: siRNA (L~25 ds)
- Viral : can be very long (L ~ 1,000,000)

Huge amounts of non-coding RNA transcribed from "junk" DNA: up to 80%

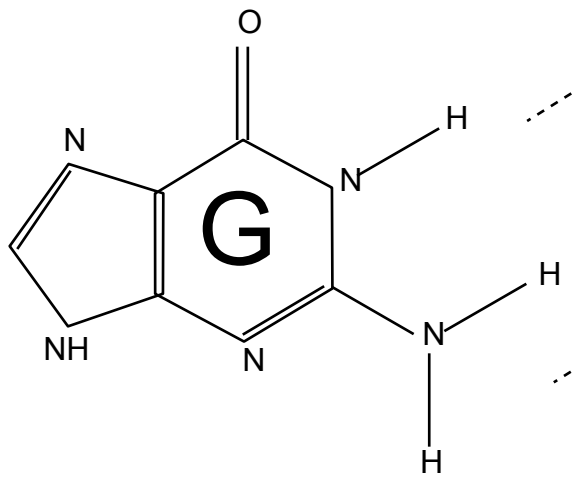
Chemistry of RNA

- 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

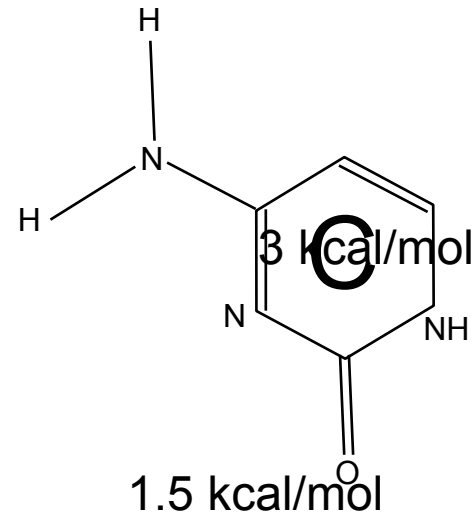
Chemistry of RNA

Guanine

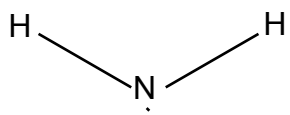


3 kcal/mol

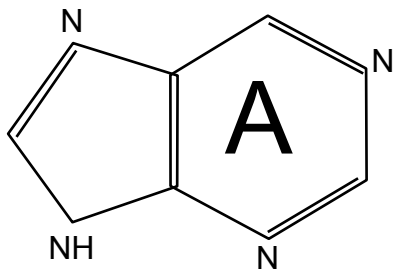
Cytosine



1.5 kcal/mol

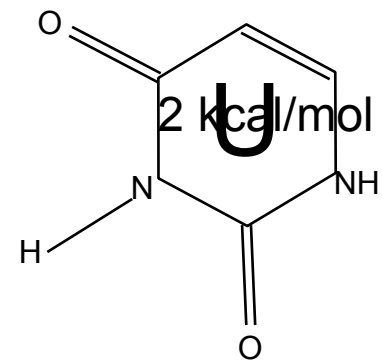


Adenine



2 kcal/mol

Uracil



Energy scales

- Crick-Watson: conjugate pairs

C – G	3kCal/mole
A – U	2kCal/mole
G – U (the wobble pair)	1.5kCal/mole

Pairings due to Hydrogen bonds between bases ⟹ RNA folding

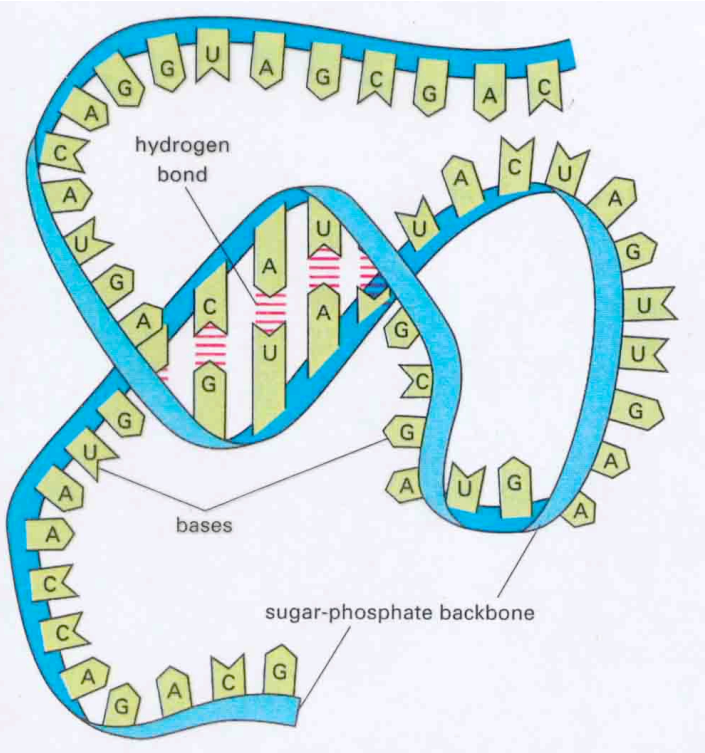
Stacking of aromatic groups

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

Base pairing

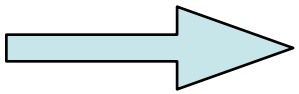
- Induces helical strands (like in DNA)
- Induces secondary structure of RNA

list of paired bases



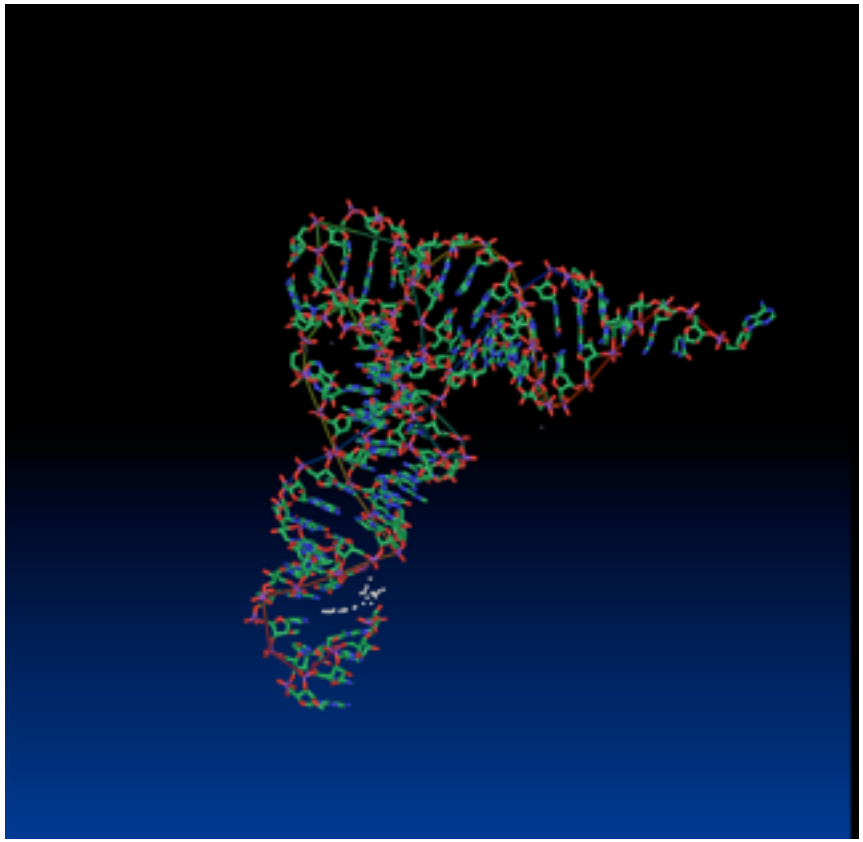
RNA folding problem: determine which bases are paired

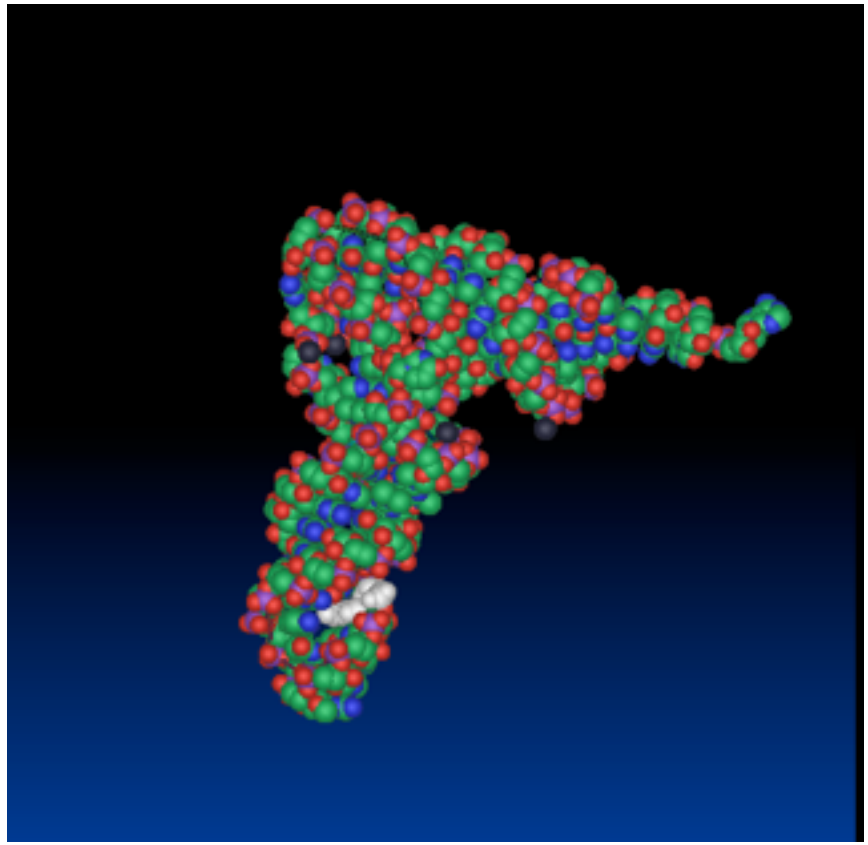
- Functions of RNA: enzyme, regulation, etc...
- strongly depends on the pairings, the loops and the pseudoknots



Must know all the pairings present in the RNA = Secondary Structure

Pictures of RNA

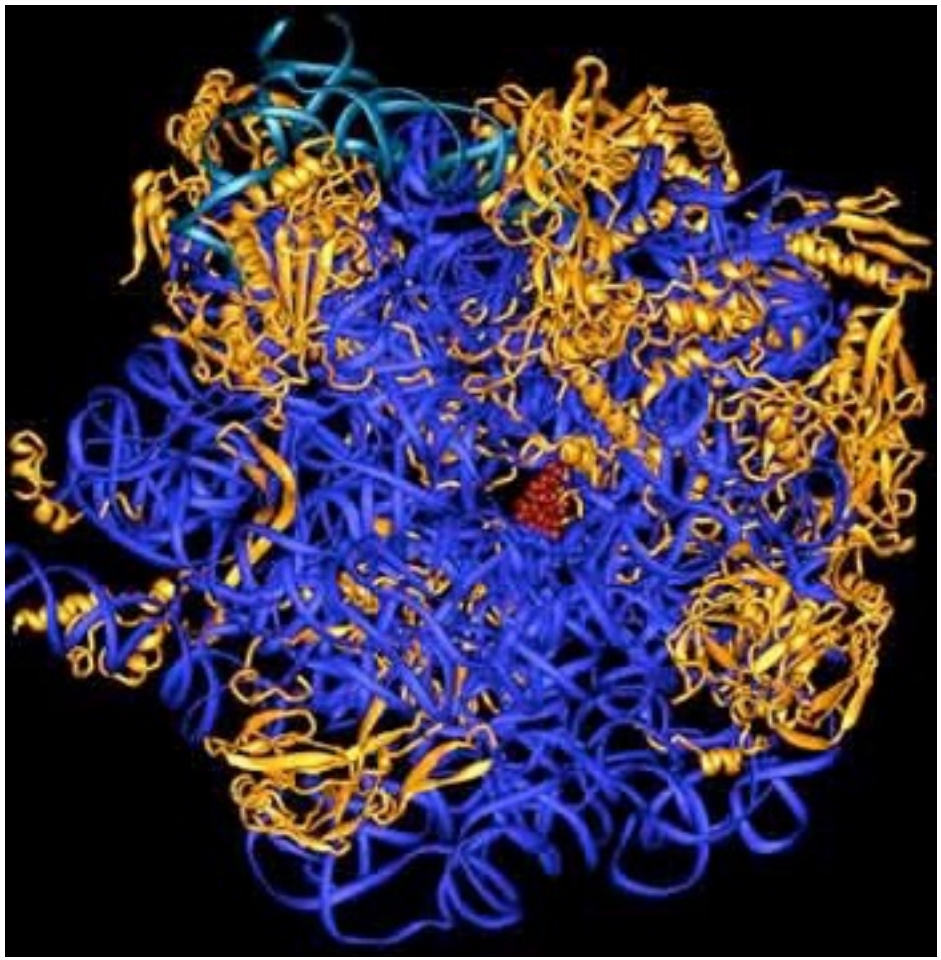


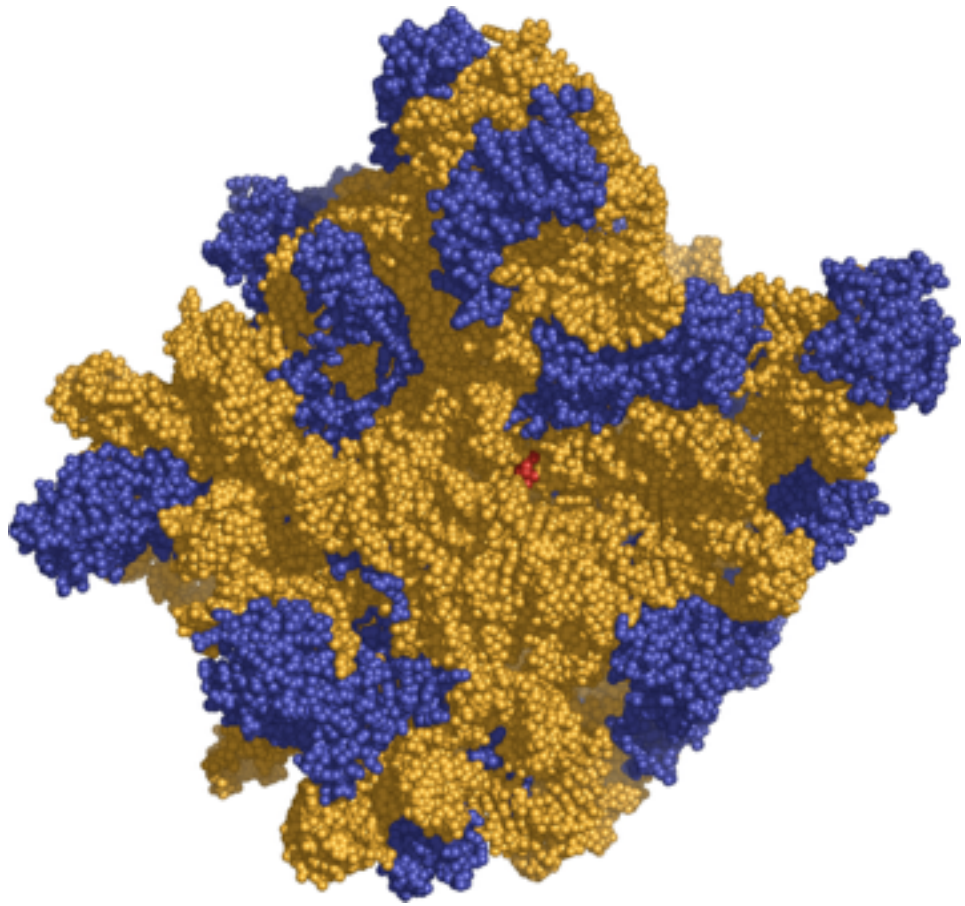


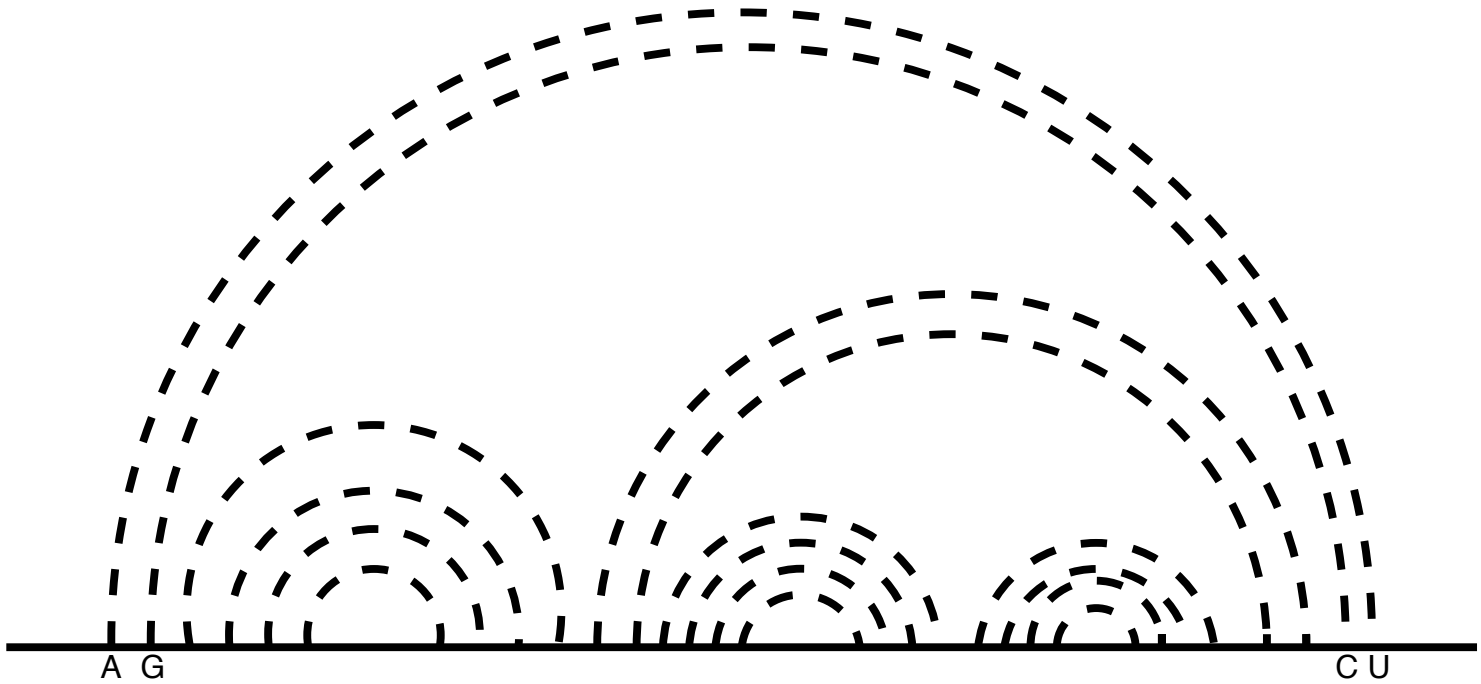
Transfer RNA

Ribosomal RNA

Nobel Prize in Chemistry 2009

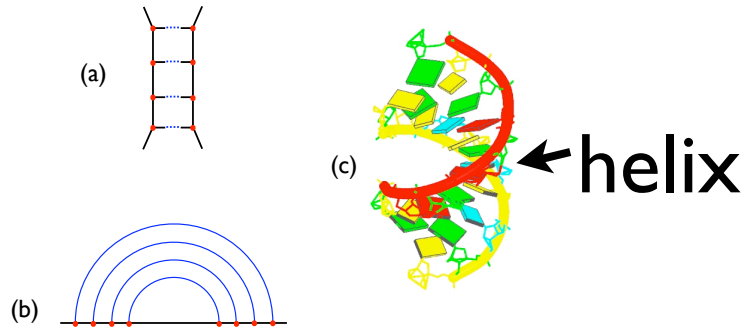


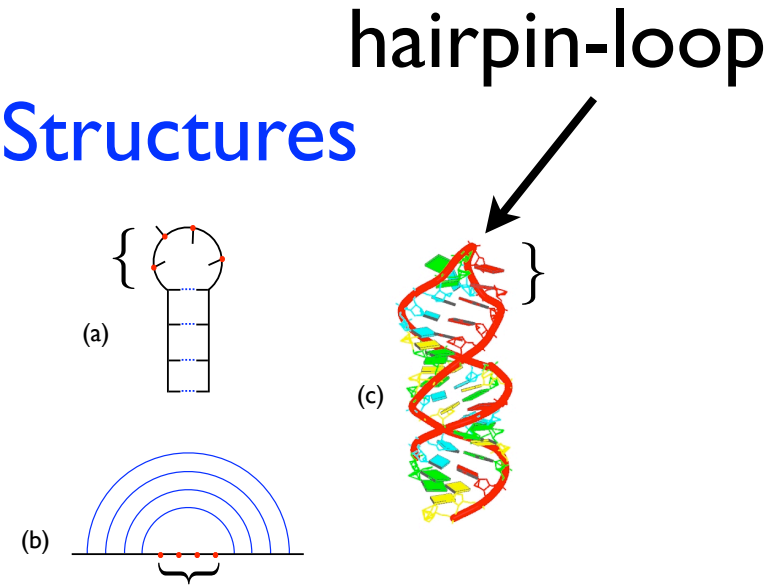


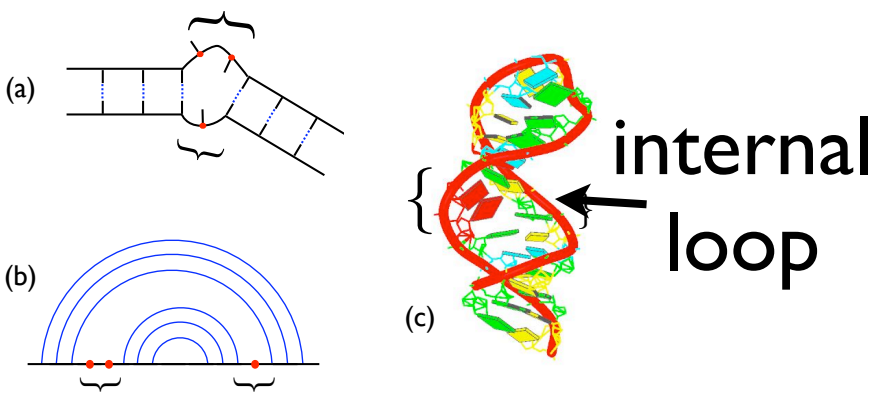


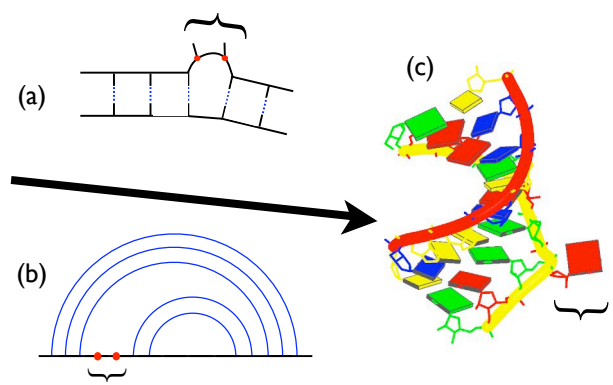
Arch representation of the secondary structure of an RNA

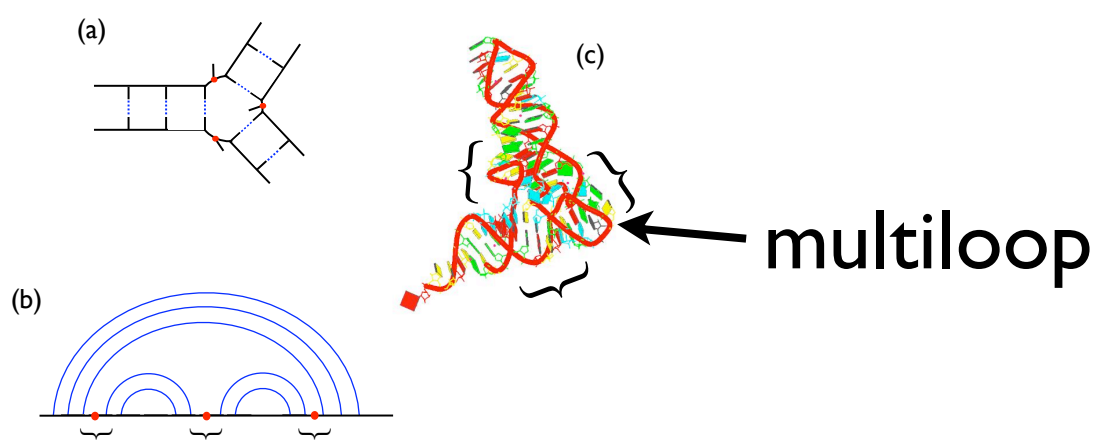
Motifs of Planar Secondary Structures



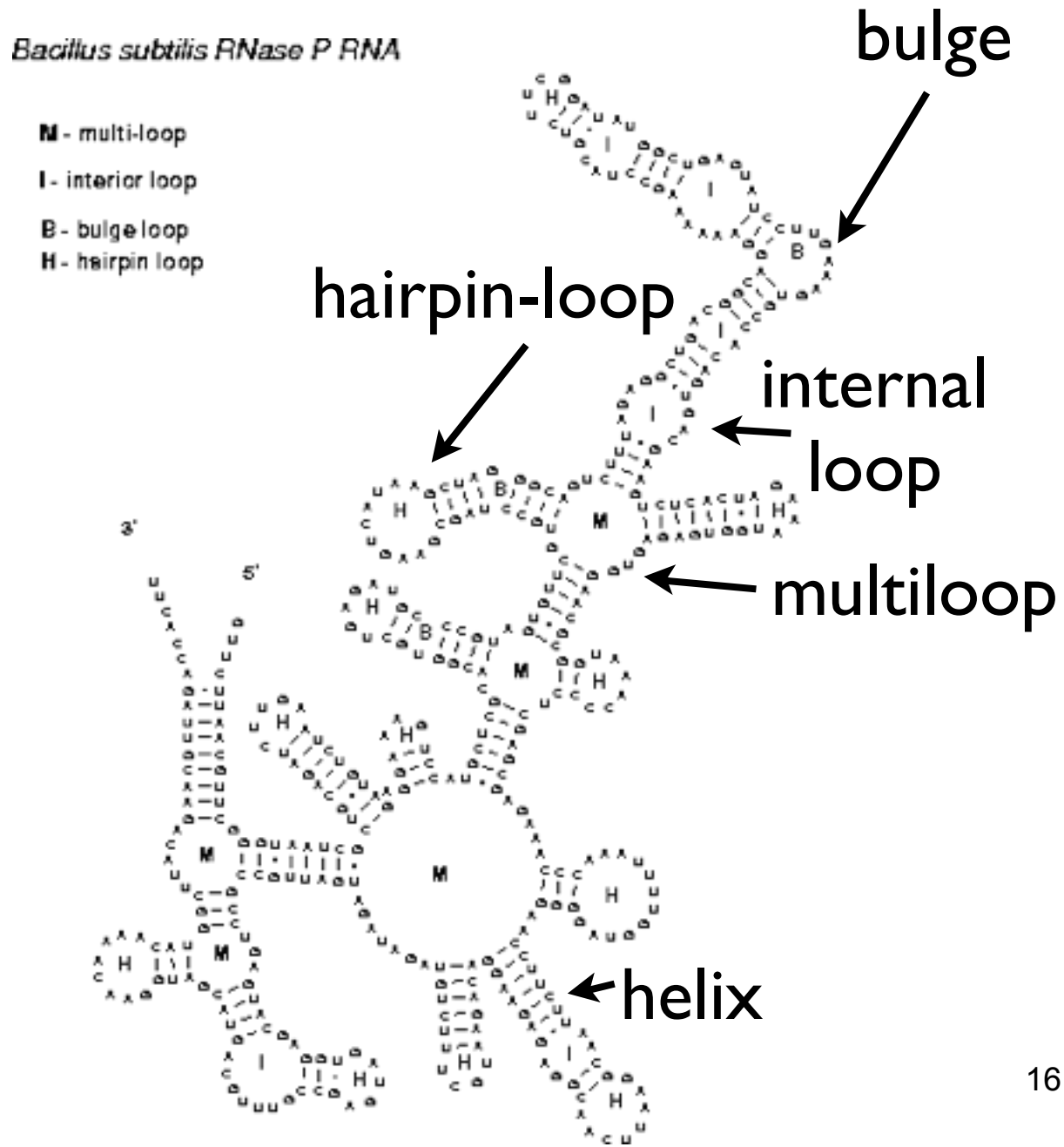






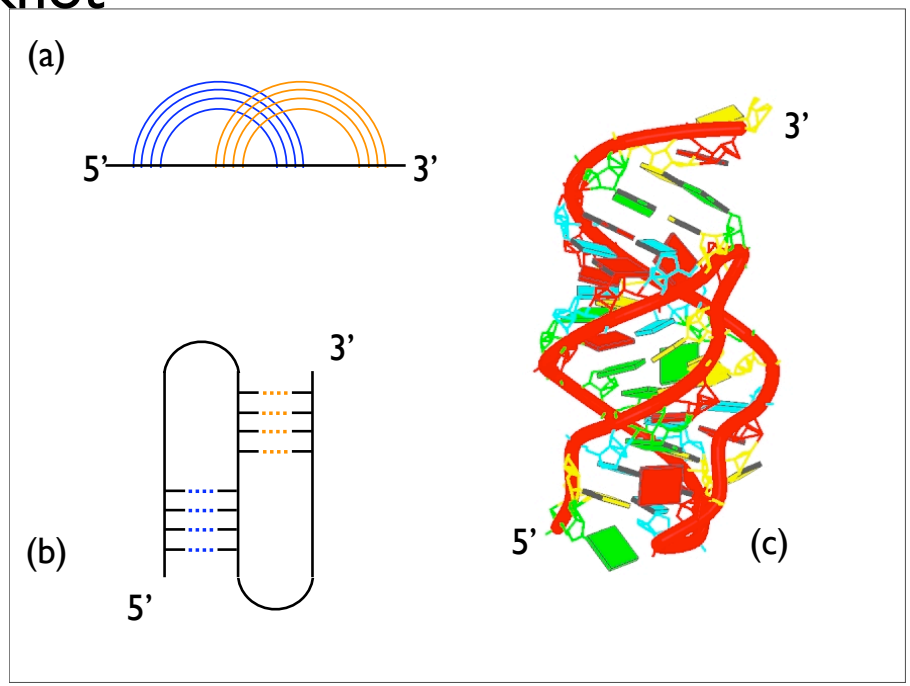


Motifs of Planar Secondary Structures



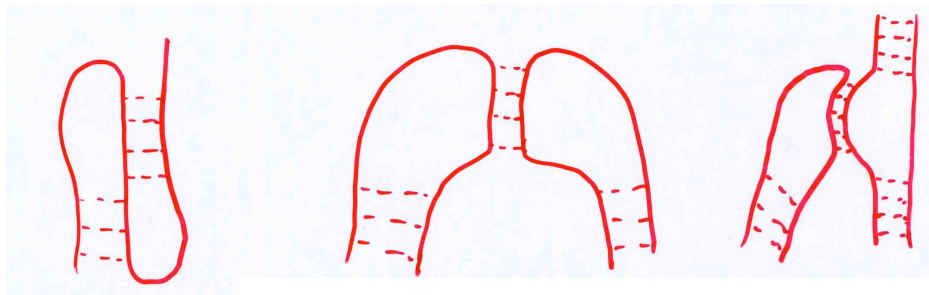
Pseudoknots

- H-Pseudoknot



the Kissing Hairpin

The H-hairpin



loop-bulge

Pseudoknots

- Small number of pseudoknots
- Less than 10% of all bases participate in pseudoknots

“Simplificity” of RNA interactions:

- Saturation of interactions
- Watson-Crick pairing

Define

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

Base pair energy

Chain rigidity

- Approximation

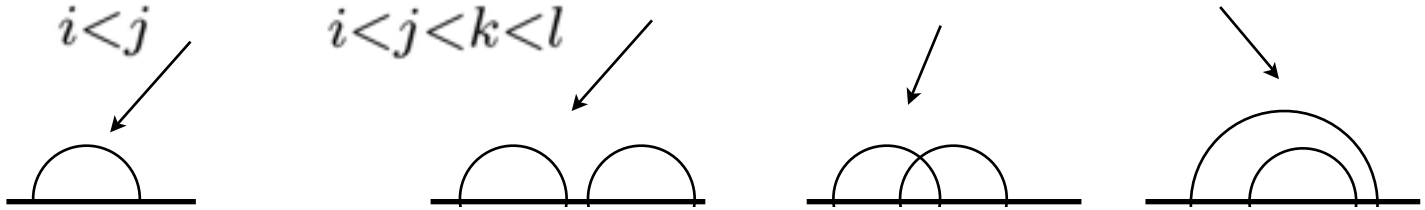
$$Z = \sum Q_0$$

sterically allowed configurations

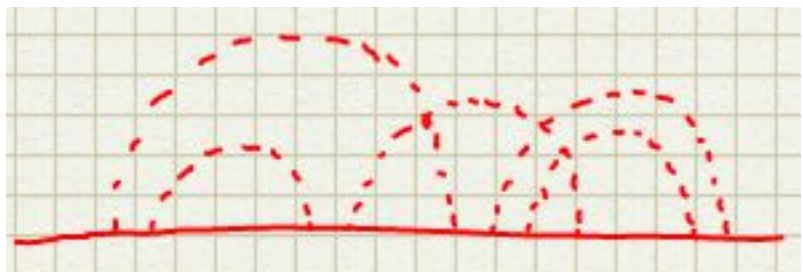
Partition Function of Secondary Structures

where

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



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

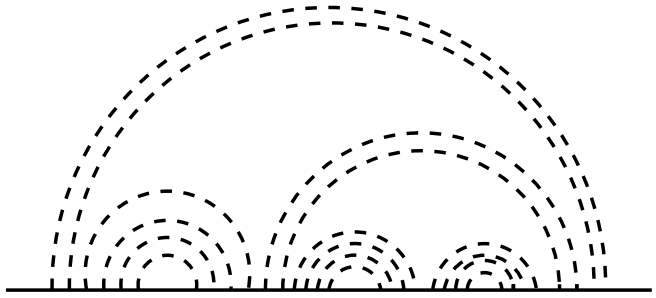


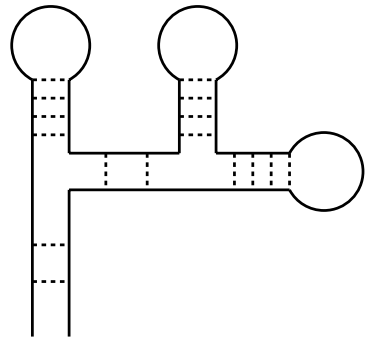
- must do the combinatorics
- any index appears once and only once (saturation)

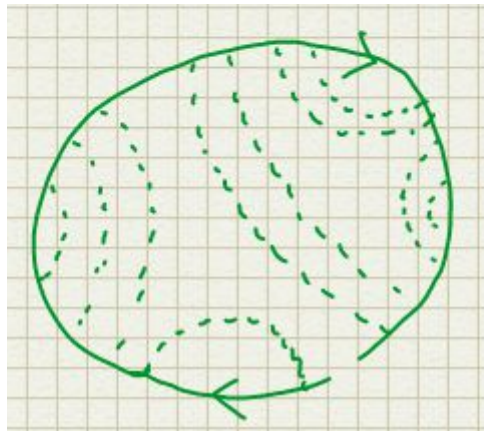
Note: analogy between pairing graphs and Feynmann graphs

Planar Secondary structures
No Pseudoknots

- We work on Q_0
- Planar Secondary structures = Arches



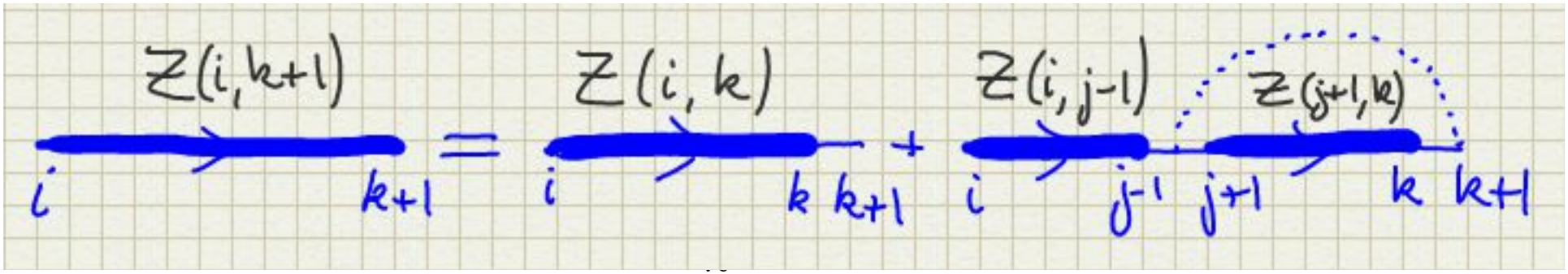




- Define $Z(i,j)$ as the
- partition function of segment (i,j)

Recursion relation

- Graphically, when one adds one base



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

- with

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

chain rigidity

- by iterating this recursion, one can generate all possible planar secondary structures, with the correct Boltzmann weights.
- Algorithm scales as N^3
- One can include Entropies and Stacking Energies
  - MFOLD
  - Vienna Package

} <60% success on tRNA

Determination of Pseudoknots is NP-complete

Wick Theorem

- Simple representation: consider an RNA sequence of length L

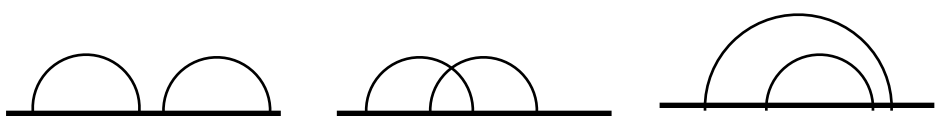
$$Q_0 = \frac{1}{\mathcal{N}} \int \prod_{i=1}^{L} d\phi_i e^{-\frac{1}{2}\sum_{i,j}\phi_i V_{ij}^{-1}\phi_j} \prod_{i=1}^{L}(1+\phi_i)$$

- 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 Pseudoknots.
- Experimentally, few pseudoknots.

Pseudoknots

- If no crossings of the arches, it is possible to calculate exactly the partition function by recursion relations: MFold, Vienna Package
- Crossings = Pseudoknots = constraints on the backbone
- Need a penalty for pseudoknots to account for mechanical constraint on backbone.

- We want to give a penalty to pseudoknots
  - which does not depend on number of crossing
  - which depends on the topological complexity of the pseudoknot
  - additive
- Matrix field theory ⟹ topology of graphs

$$Z(1,L) = \frac{1}{A(L)} \int \prod_{k=1}^{L} \prod_{a \le b} d\phi_k(a,b) e^{-\frac{N}{2} \sum_{i,j} (V^{-1})_{ij} \mathrm{Tr} \phi_i \phi_j} \frac{1}{N} \mathrm{Tr} \prod_{k=1}^{L} (1+\phi_k)$$

where $\phi_k(a,b)$ is an $N \times N$ real symmetric matrix

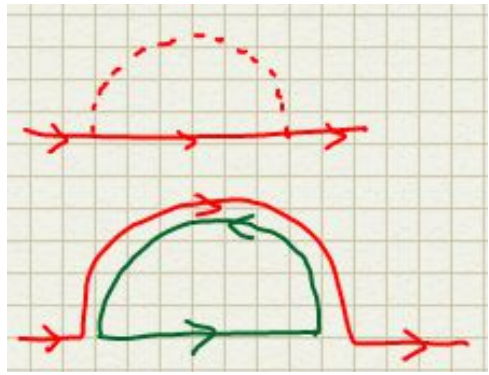
$$Q_0 = \frac{1}{\mathcal{N}} \int \prod_{i=1}^{L} d\phi_i e^{-\frac{1}{2} \sum_{i,j} \phi_i V_{ij}^{-1} \phi_j} \prod_{i=1}^{L} (1+\phi_i)$$

Double line graphs

G.t'Hooft (1973)

- Matrix fields ⟹ double line graphs

$\phi_i(a,b)$: *NxN* matrix



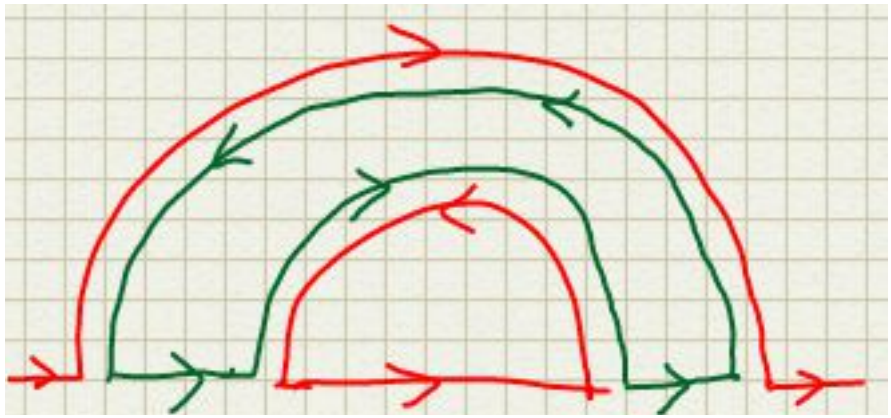
- If we use the rule:

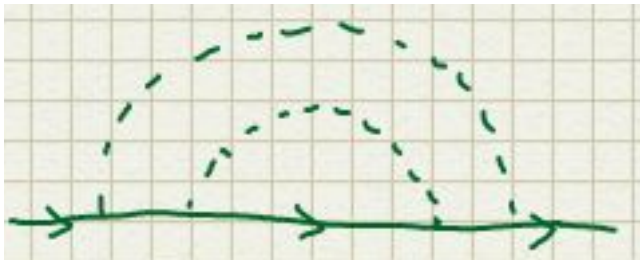
Propagator: *1/N*

Loop: *N*

- Above graph: ⟹ $N \times \frac{1}{N} = 1$

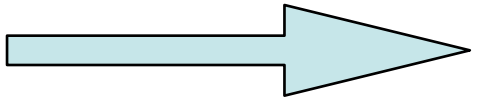
- Other graph





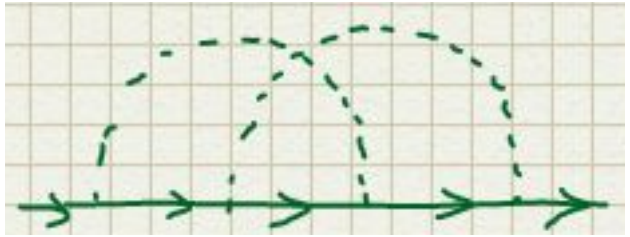
2 internal lines: $1/N^2$

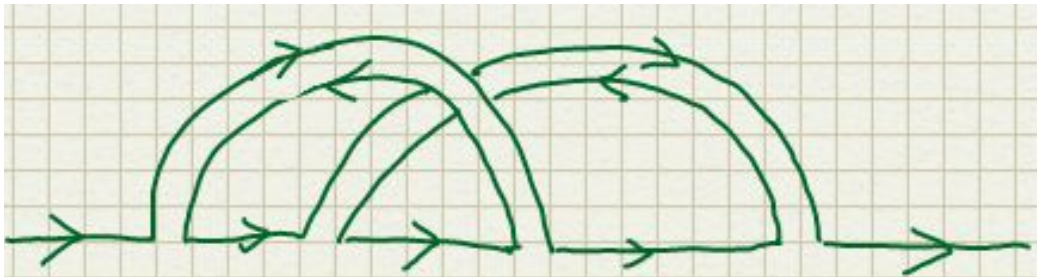
2 Loops: N^2



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*

- By looking at a few diagrams, Matrix Field Theory seems to do what we want:
  - Hartree (Planar) diagrams are of order 1
  - Pseudoknots are of higher order in 1/N

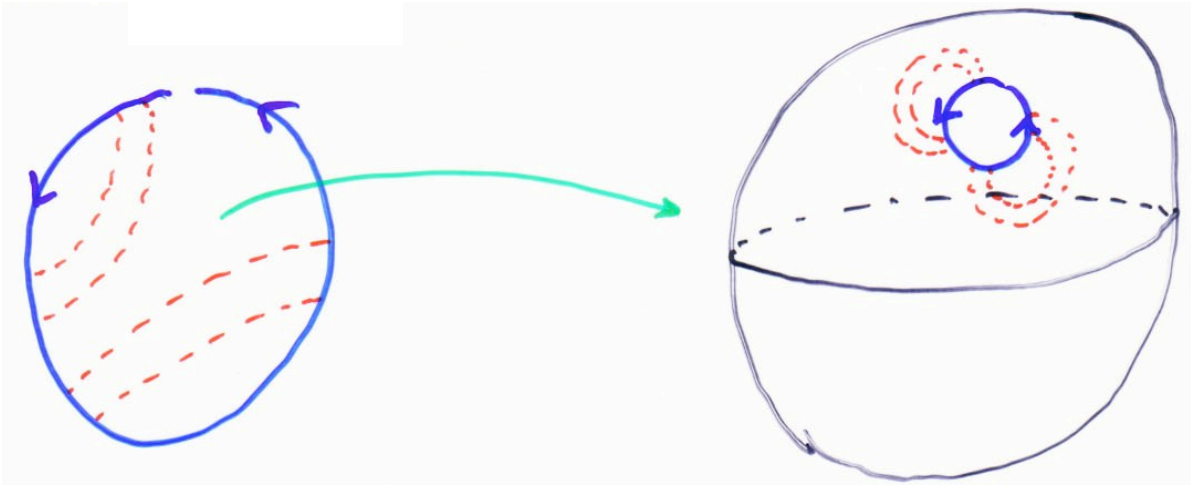
- One can prove that the matrix field partition function is equal to

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

- where *g*(pairing) is the genus of the pairing graph
- each graph of the matrix theory carries a Boltzmann factor and is weighted by a factor $\frac{1}{N^{2g}}$

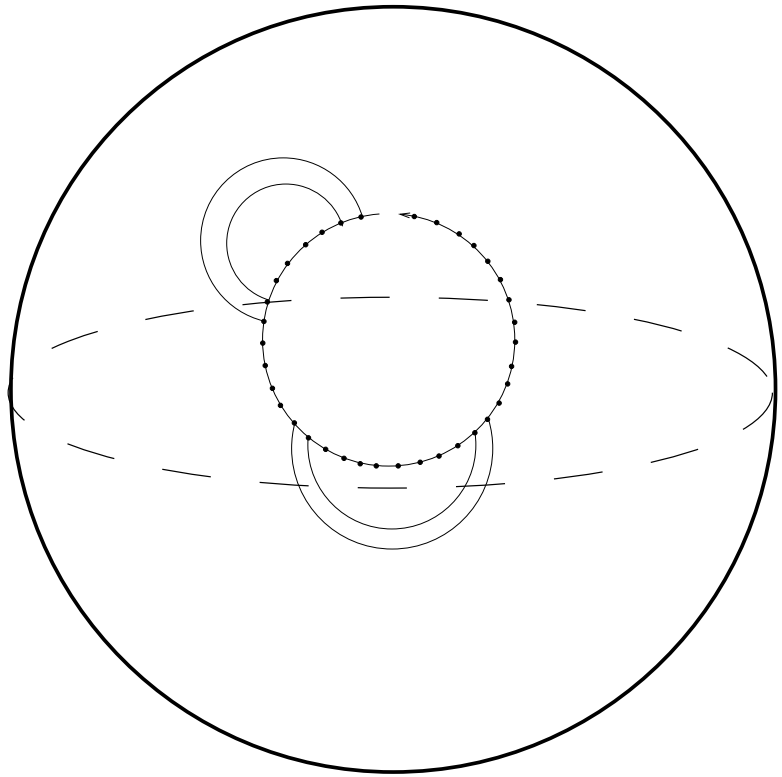
Topological classification of RNA folds

- An RNA fold can be characterized by its topology:

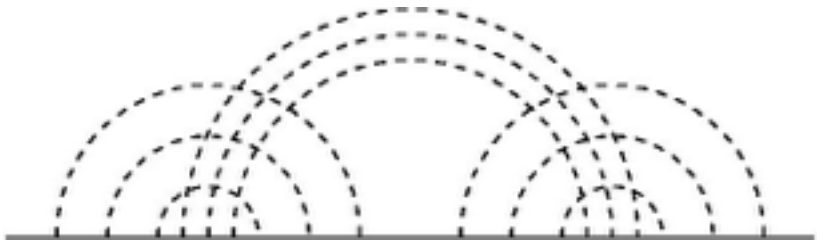


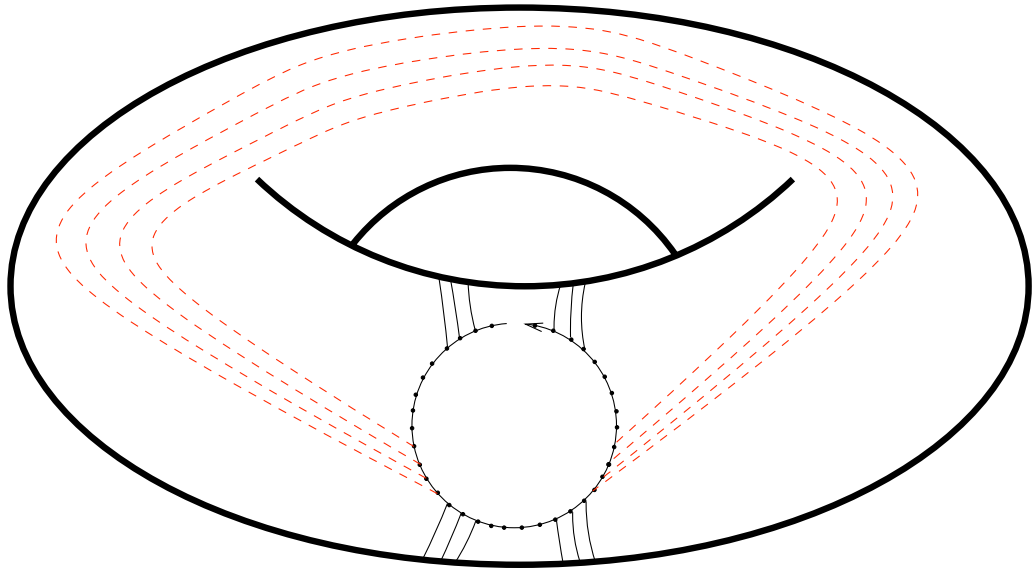
- Genus: Minimum number of handles of embedding surface

Genus 0: the Sphere

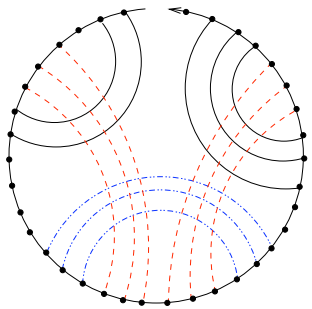


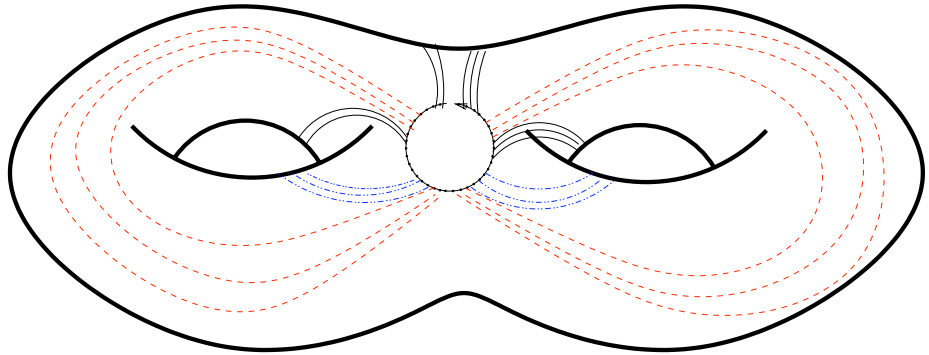
Genus 1: the Torus

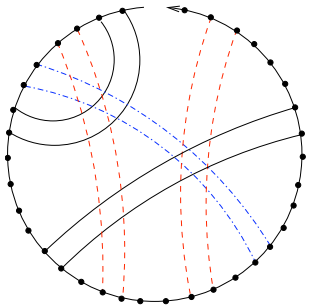


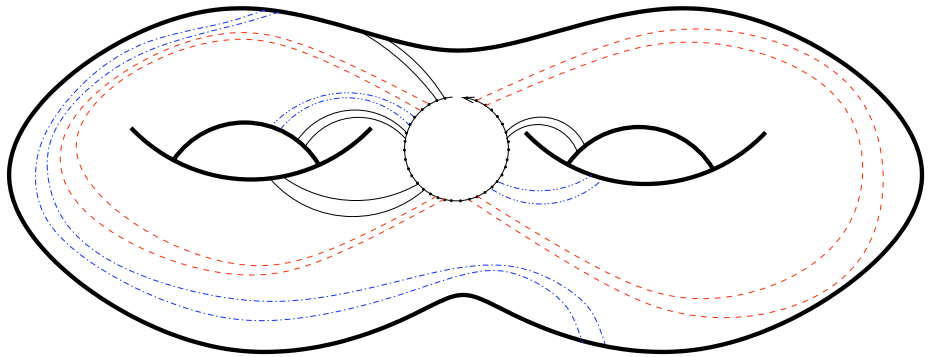


Genus 2: the Bi-torus

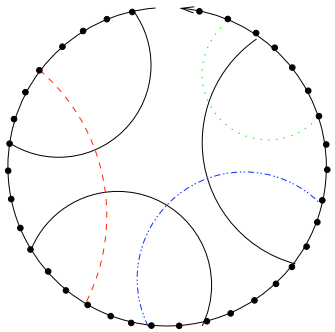


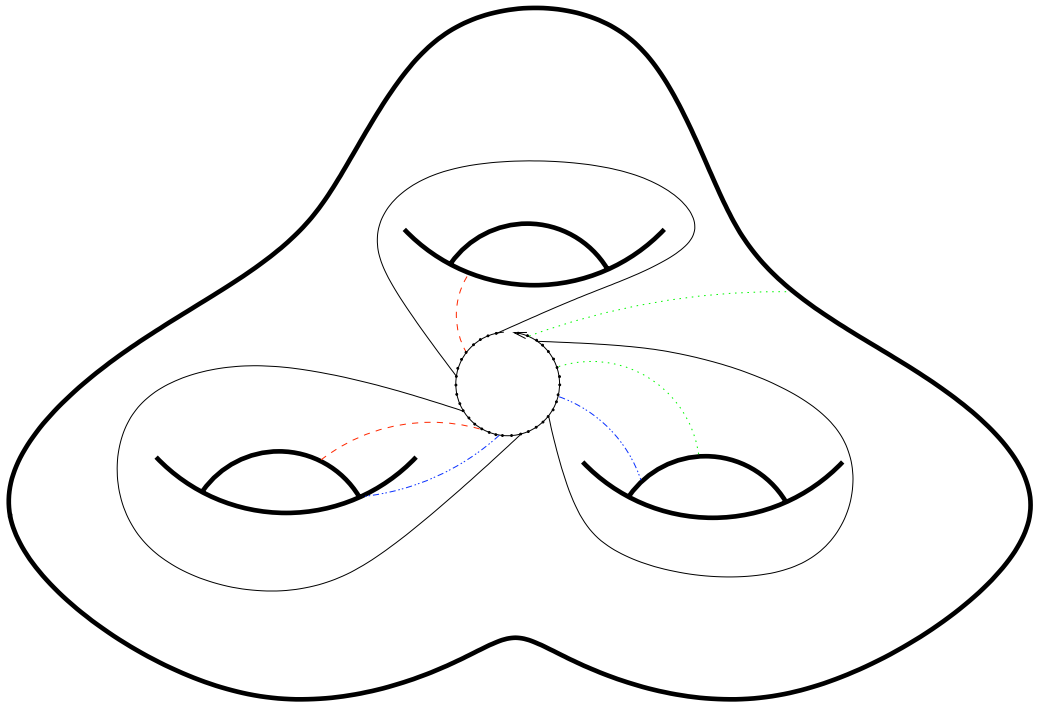






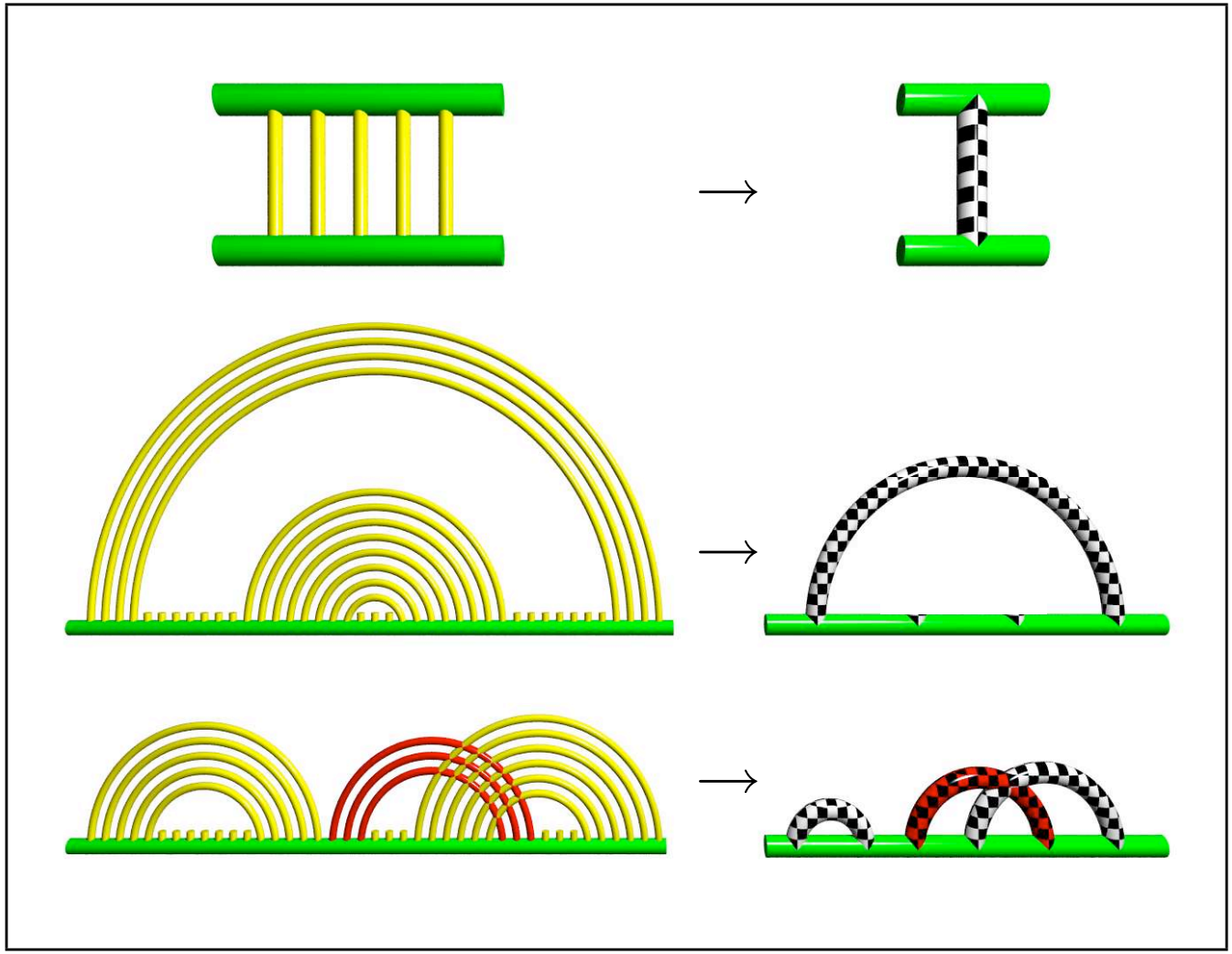
Genus 3



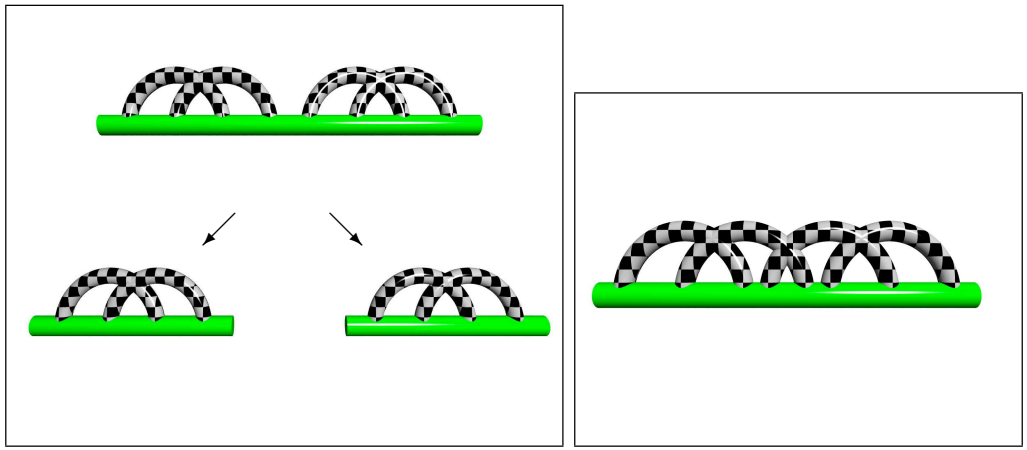


Graphology

Parallel pairings don’t change the genus

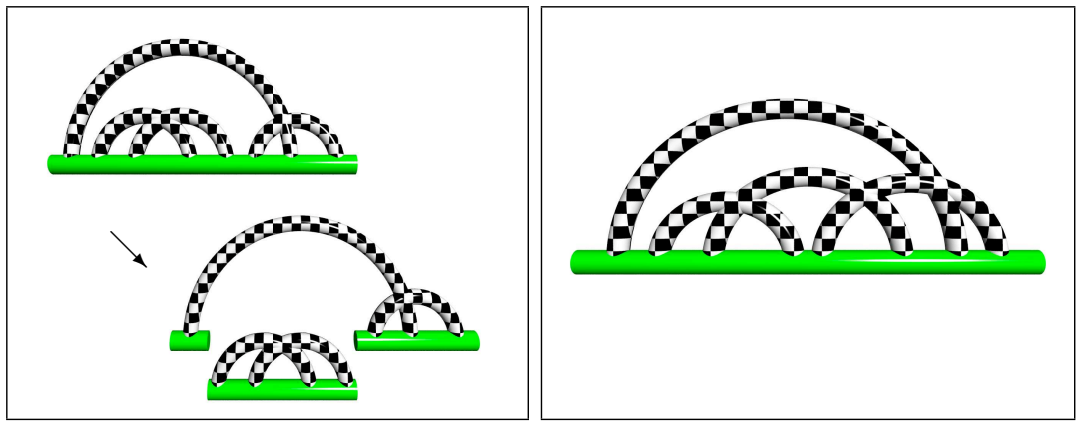


Irreducibility and Nesting



Irreducible PK

Genus is additive

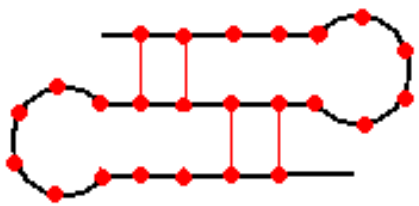


Non nested PK

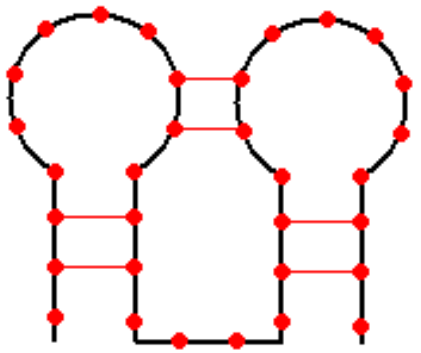
Only 4 primitive PK of genus 1

Primitive=Irreducible and non-nested

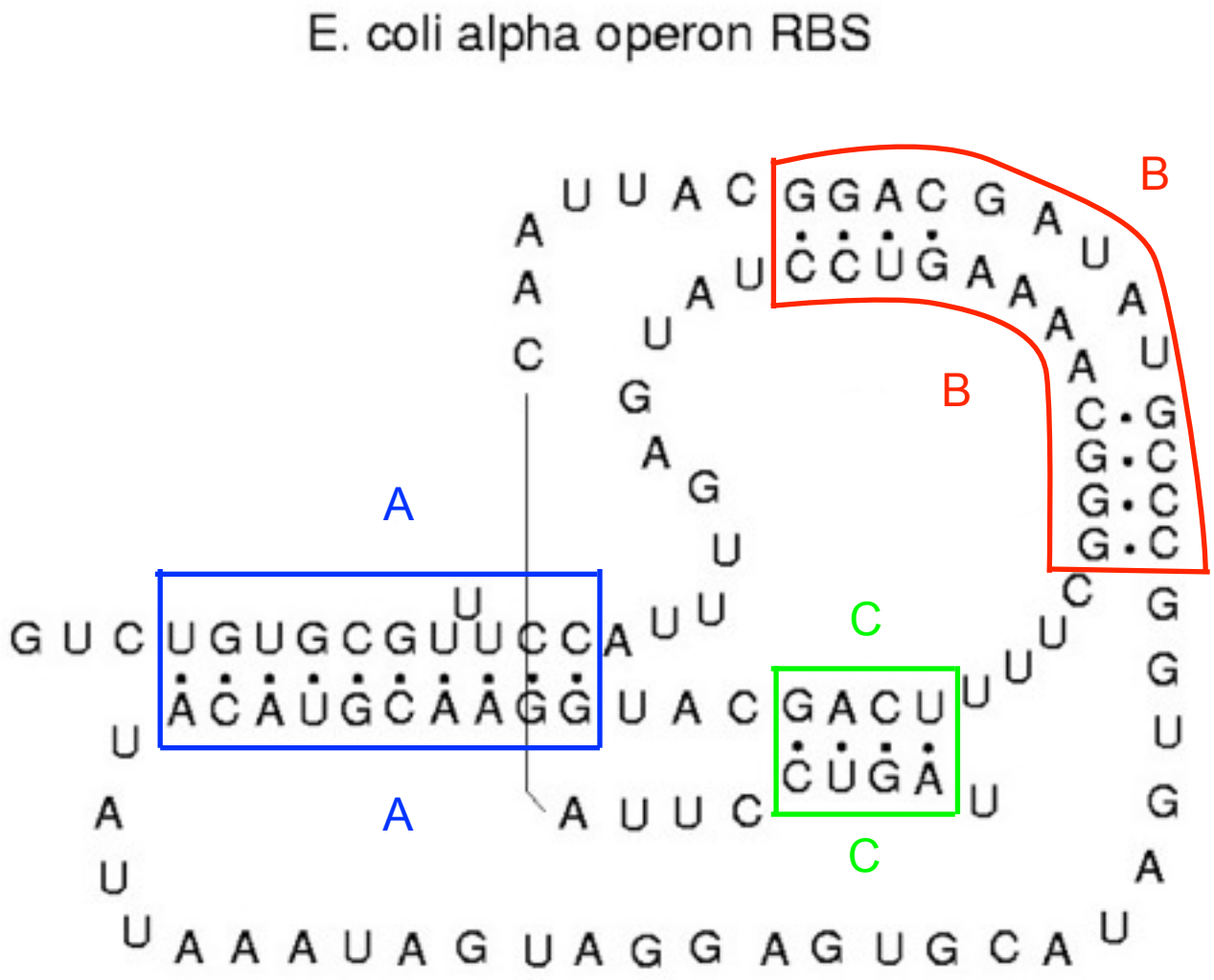
H PK



KH P



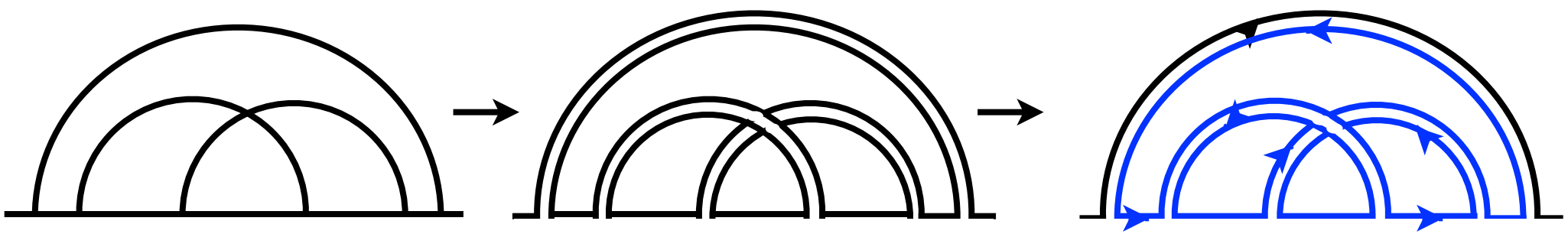
An exemple of ABCABC pseudo-knot



How to compute the genus?

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

$$g = \frac{2 - 2}{2} = 0$$



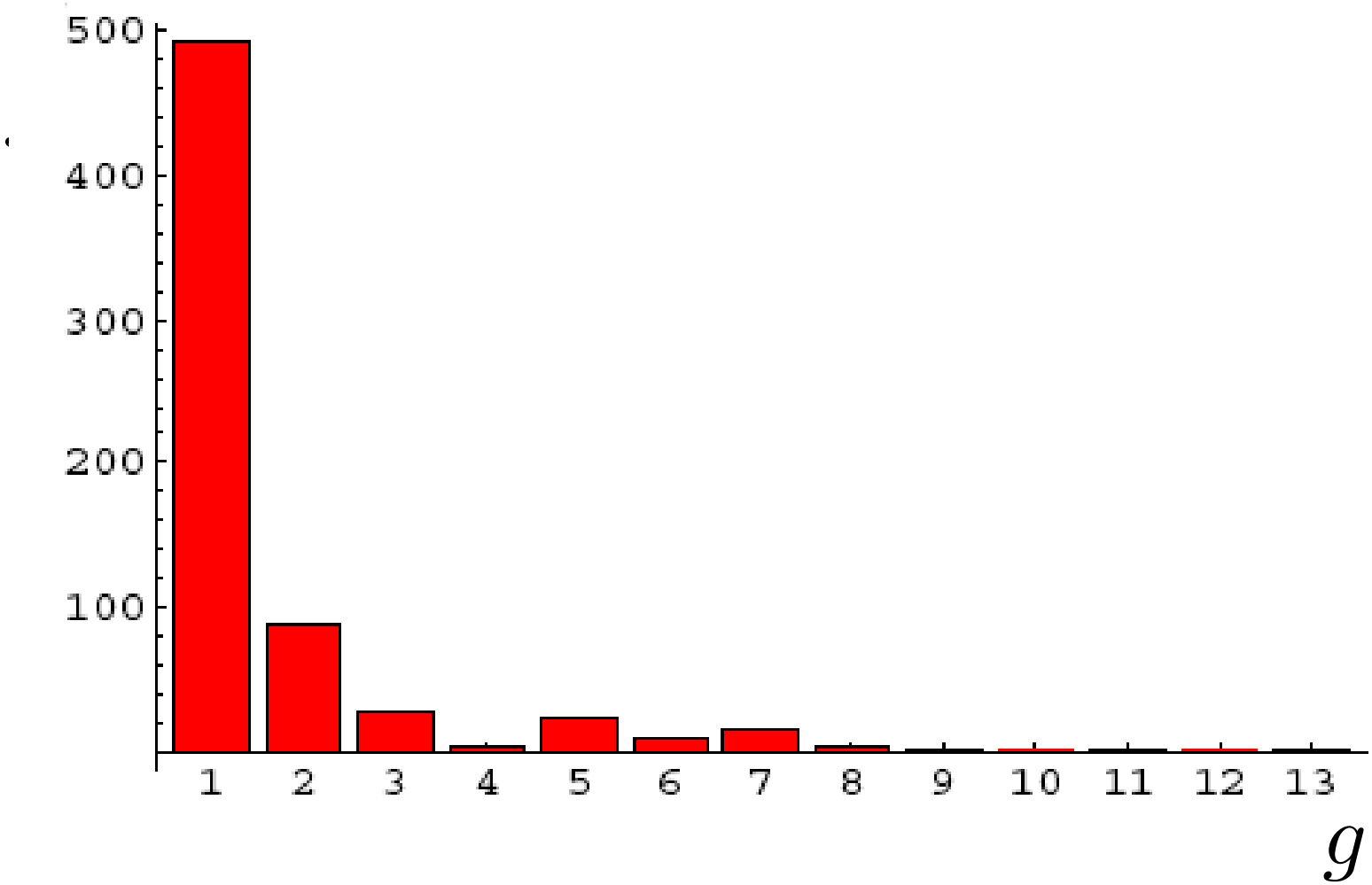
$$g = \frac{3 - 1}{2} = 1$$

- Protein Data Bank (PDB): 1025 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=75) are KHP of genus 1

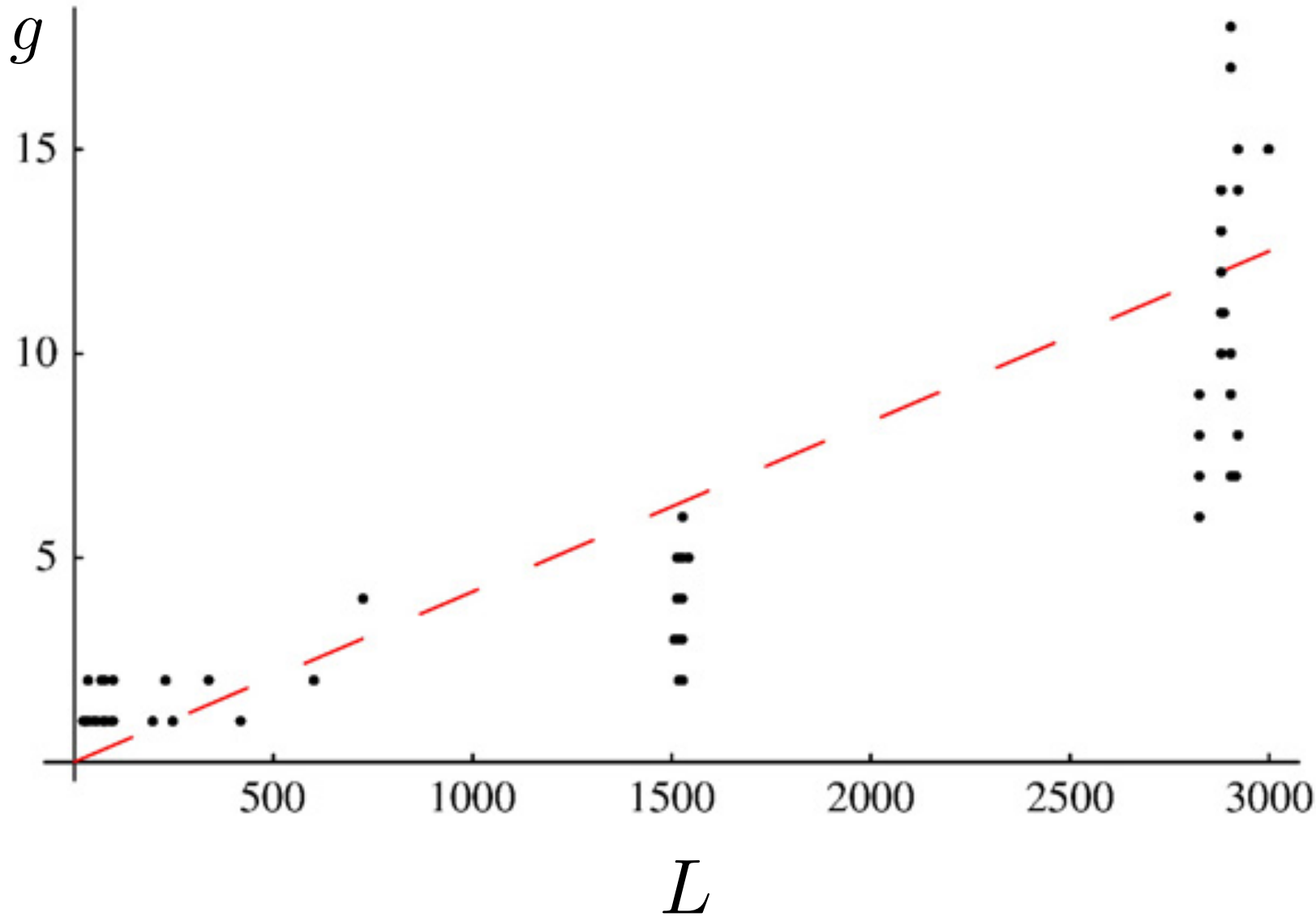
Statistical study

- 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

Histogram of the number of RNA as a function of the genus



Genus as a function of sequence length



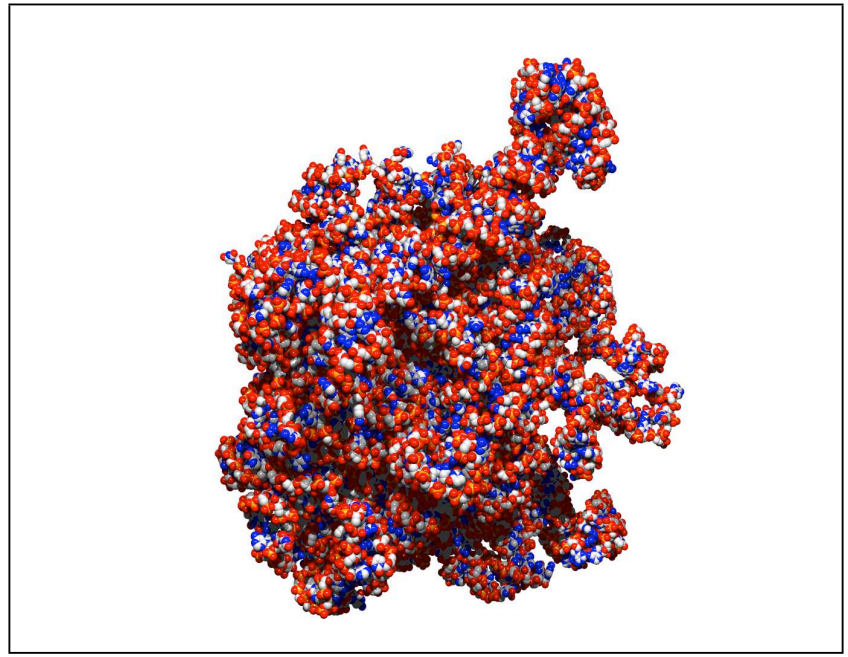
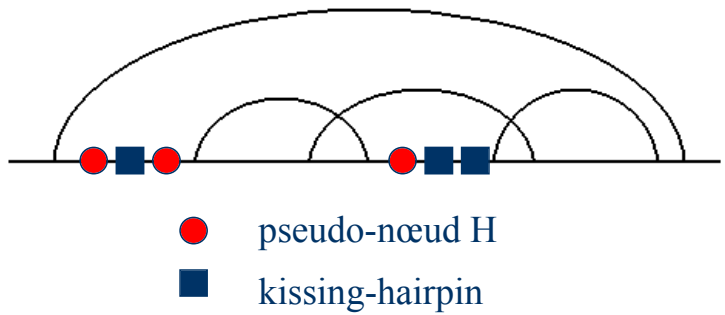


Figure 11: The B chain of 1vou.pdb is an RNA of genus 7 and of length 2825 bases.

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



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 a one matrix integral:

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

- 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

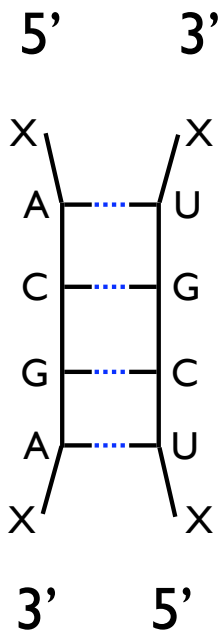
$$< g >_L \approx 0.25L$$

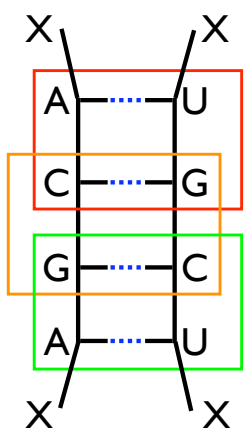
- for real RNA, the largest genus we found is 18 for ribosomes (size around 3000 bp). The genus should be around 750.
- if one includes self-avoidance of chain, we find

$$< g >_L \approx 0.13L$$

Free Energy Parametrization

- Stacking free energies
- Penalty for loop opening
- Penalty for bulges
- No Conformational Entropy
- Penalty proportional to the genus of PK: μg





$$\Delta F(\text{hélice}) = \Delta F_t \begin{pmatrix} \text{X} & \text{X} \\ \text{A}-\text{U} \end{pmatrix} + \Delta F_d \begin{pmatrix} \text{A}-\text{U} \\ \text{C}-\text{G} \end{pmatrix} + \Delta F_d \begin{pmatrix} \text{C}-\text{G} \\ \text{G}-\text{C} \end{pmatrix} + \Delta F_d \begin{pmatrix} \text{G}-\text{C} \\ \text{A}-\text{U} \end{pmatrix} + \Delta F_t \begin{pmatrix} \text{A}-\text{U} \\ \text{X} & \text{X} \end{pmatrix}$$

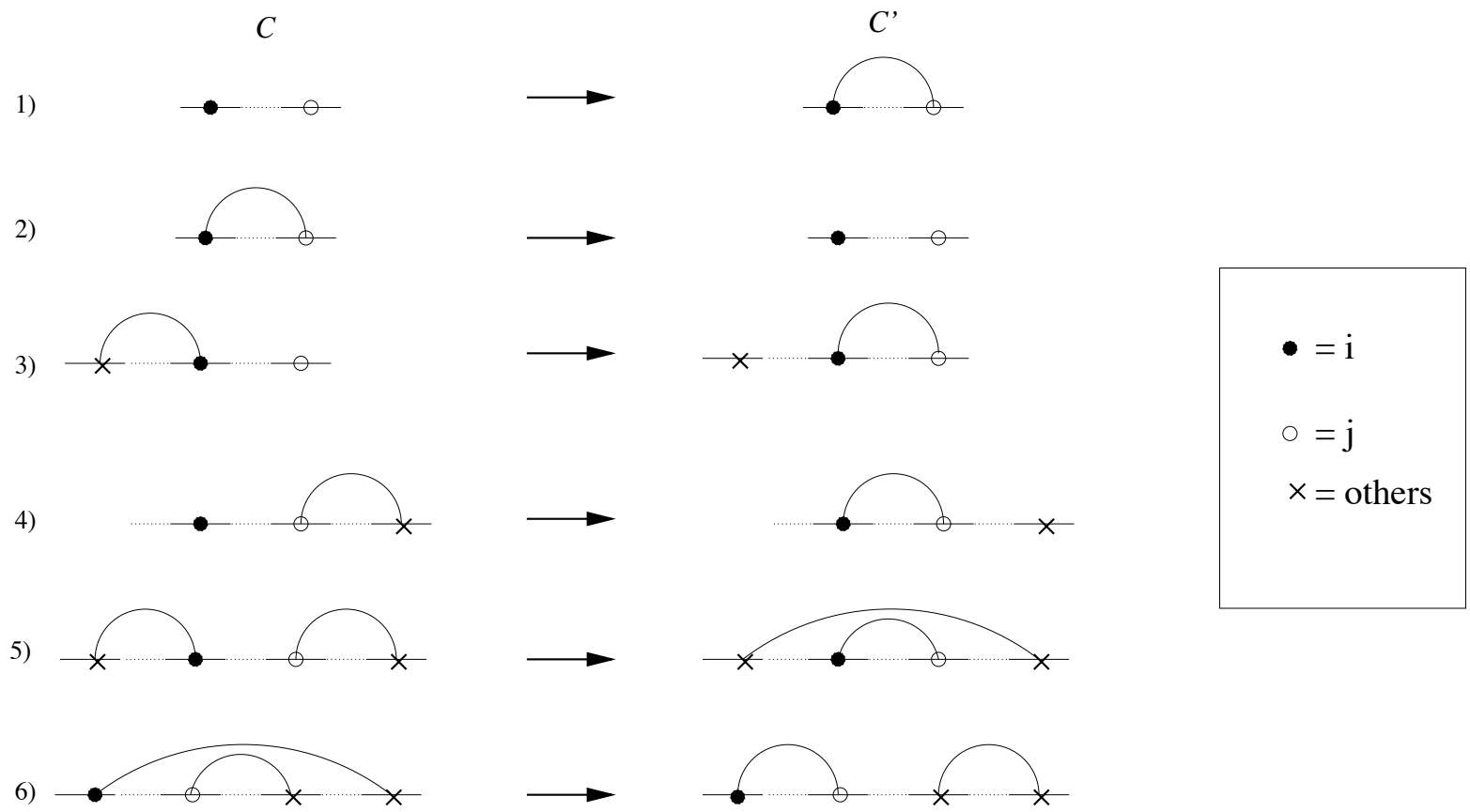
Monte Carlo Method

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

where

$$e^{-\mu} = \frac{1}{N^2}$$

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

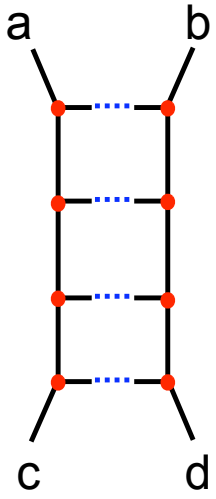
$$p = e^{-\beta \Delta E - \mu \Delta g}$$

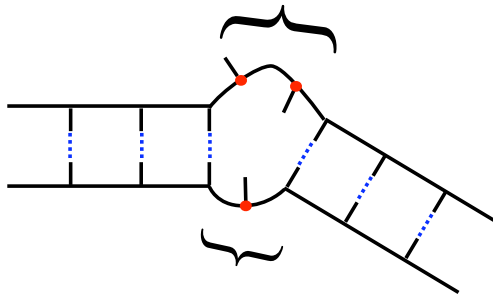
- But Monte Carlo is not a good method for helices: high barriers to open a helix
- Must use another algorithm for calculation of Free Energy

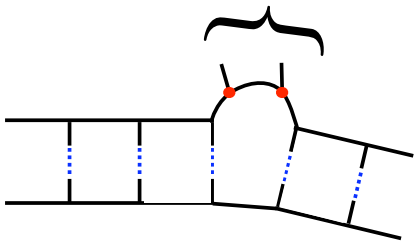
TT2NE: Structure Building

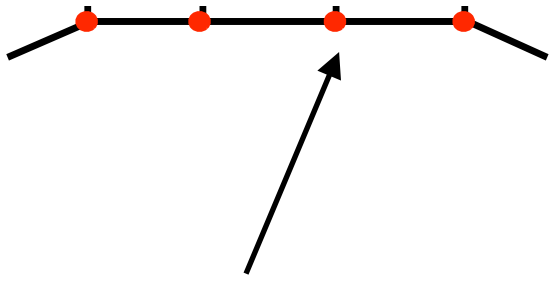
- Build a library of all possible favorable fragments of helices (negative free energies). The rest are unpaired segments.
- Put as many helices as possible in a graph and join them by unpaired segments. Compute the genus and the free energy.
- 2 helices are incompatible if
  - they share common bases
  - their concatenation produces an existing larger helix
  - they produce a sterically impossible structure

LIBRARY









complete with

TT2NE

http://ipht.cea.fr/rna/tt2ne.php

- Minimum free energy structures can be obtained by:
  - Exact enumeration for *L<150*
  - Heuristic (limited depth graph exploration) for *L<250*

McGenus

http://ipht.cea.fr/rna/mcgenus.php

- Do the same but with simulated tempering: run systems at several temperatures in parallel, and exchange systems at different temperatures.
- Add or remove helix stochastically and accept or reject with Metropolis scheme
- Possibility to give penalty for each topolgy
- Works for sizes up to 1000 bases: tmRNA, etc...

Results on a test database of 50 RNA with pseudoknots $l < 229$

sequence	exp	l	Mfold	HotKnots	McQfold	ProbKnot	TT2NE	g	g_T	g_{HK}	g_{MQ}	g_{PK}
1u8d	X	68	69 - 100	69 - 100	69 - 100	69 - 100	**88** - 100	1	1	0	0	0
1y0q	X	229	65 - 75	63 - 75	66 - 71	**75** - 91	**75** - 70	1	2	0	1	0
AMV3	X	113	84 - 86	84 - 86	76 - 81	84 - 80	**87** - 85	1	1	0	0	0
BBMV		116	81 - 81	81 - 81	**86** - 82	84 - 82	**86** - 84	1	1	0	1	0
BMV3	X	138	84 - 86	27 - 66	27 - 70	84 - 80	**100** - 97	1	1	0	0	0
Bp_PK2		91	81 - 96	81 - 96	87 - 87	81 - 81	**100** - 100	1	1	0	1	0
Bs_glms	X	158	42 - 43	44 - 46	**76** - 85	**76** - 83	65 - 57	2	3	0	0	0
BVDV		74	52 - 65	52 - 61	76 - 82	48 - 57	**96** - 96	1	1	0	1	0
BWYV	X	51	55 - 55	**100** - 69	55* - 55	55 - 100	**100** - 100	1	1	1	1	0
SRV-1	X	38	0 - 0	**100** - 100	**100** - 100	0 - 0	**100** - 100	1	1	1	1	0
TEV	X	94	21 - 31	21 - 31	28 - 53	32 - 60	28 - 47	3	0	0	0	0
T2_gene32	X	33	58 - 70	**100** - 100	**100** - 100	58 - 70	**100** - 100	1	1	1	1	0
T4_gene32	X	28	63 - 87	63* - 87	63 - 100	63 - 100	**100** - 100	1	1	0	0	0
TMV	X	74	52 - 65	52 - 61	52 - 65	**56** - 60	48 - 54	3	1	0	0	0
Tt-LSU	X	65	60 - 75	**95** - 100	60- 100	60 - 80	**95** - 100	1	1	1	0	0
TYMV	X	74	**72** - 78	70 - 73	**72** - 78	**72** - 78	**72** - 69	1	1	0	0	0
VMV	X	69	50 - 41	50 - 41	**100** - 60	50 - 35	**100** - 70	1	1	0	1	0
average1			55 - 60	60 - 67	66 - 75	60 - 64	**78** - 76					
average2			50 - 58	63 - 67	68 - 77	54 - 63	**80** - 78					
st-dev			26 - 32	32 - 32	29 - 25	24 - 30	30 - 28					

sensitivity= number of correctly predicted pairs/ number of pairs in the real structure

PPV (positive predicted value) = number of correctly predicted pairs/number of pairs of the predicted structure

public server at
http://ipht.cea.fr/rna/mcgenus.php

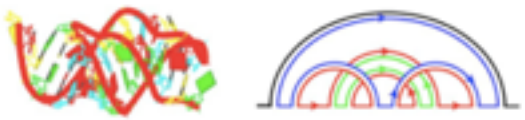
- For 590 sequences of tmRNA) (200 < *l* < 500), all previous methods yield sensitivity < 43% while McGenus yields 58%.
- A representative set of sequences of size between 200 and 300 achieve around 80% sensitivity.

- In all cases, errors can be traced back to steric constraints.

http://ipht.cea.fr/rna/mcgenus.php

McGenus & TT2NE

Algorithms for RNA pseudoknot prediction



Folding an RNA with pseudoknots

RNA sequence

Write or paste an RNA sequence : list of bases A, C, G, U (upper or lowercase) with or without blanks between the letters.
The server can only treat sequences of length smaller than 1000.

For sequences smaller than 500 bases, the computation time is less than 10 minutes.
For longer sequences, the computation time may reach one hour.
To obtain the executable of McGenus, please contact: michael.bon@cea.fr or michele@sissa.it or henri.orland@cea.fr .

Parameters

Maximum genus : 3
Genus penalty : 1.5
Number of optimal structures : 10

Reset

Fold it !

McGenus !

McGenus constrained ! to enforce only H-pseudoknots of extent less than 70 bases

TT2NE !

Fold it !

McGenus !

McGenus constrained ! to enforce only H-pseudoknots of extent less than 70 bases

TT2NE !

OR

Calculation of the genus of a structure

Upload a structure file in bpseq or ct format

Choose File no file selected

What is the format of your uploaded file ?

○ bpseq ◉ ct

Genus !

About the algorithms

The McGenus and TT2NE algorithms output predictions of RNA secondary structures with pseudoknots, based on penalizing or restricting the topological genus of the pairing graphs. The topological genus is an indicator of the complexity of the topology of the pairing [1].

The McGenus algorithm performs a stochastic Monte Carlo search in pairing space for sequences of up to 1000 bases [2].

The TT2NE algorithm performs an exhaustive or partially exhaustive search in pairing space for sequences of up to respectively 100 or 225 bases [3].

To obtain the executable of McGenus, please contact: michael.bon@cea.fr or michelet@sissa.it or henri.orland@cea.fr .

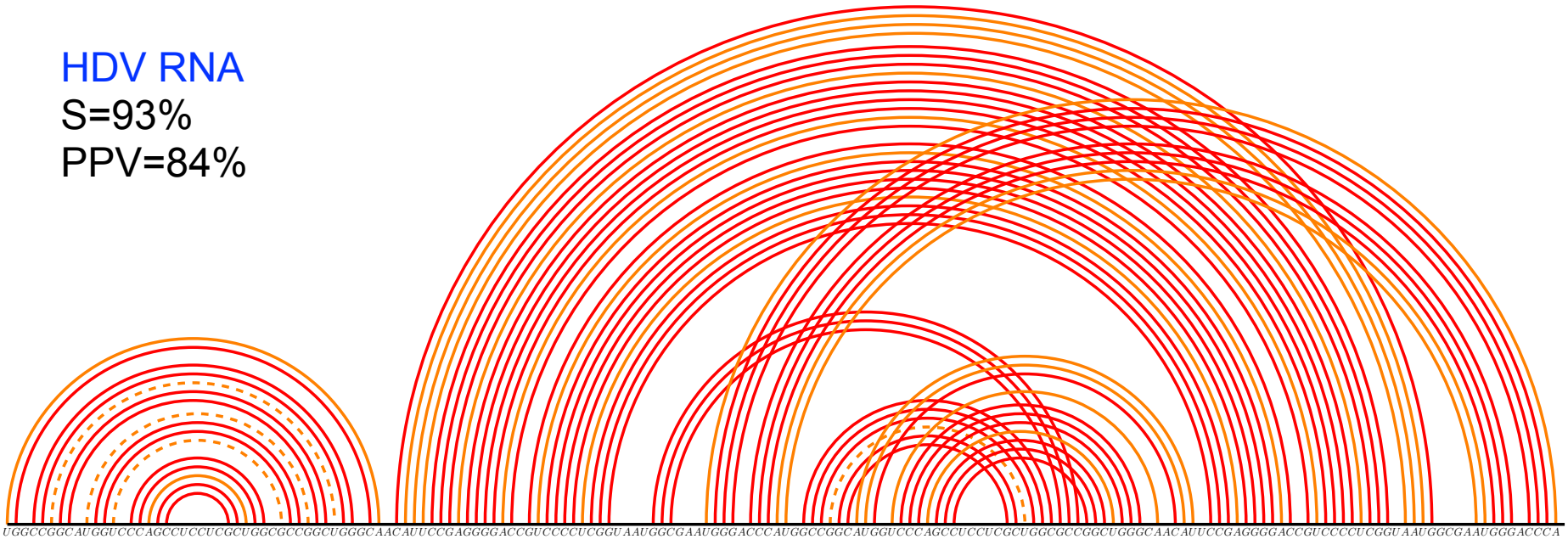
References

1. M. Bon and H. Orland, *Prediction of RNA secondary structures with pseudoknots*, Physica A (2010). Link to article
2. M. Bon, C. Micheletti and H. Orland, *McGenus: A Monte Carlo algorithm to predict RNA secondary structures with pseudoknots*, Nucleic Acids Research (2012) Link to article
3. M. Bon and H. Orland, *TT2NE: A novel algorithm to predict RNA secondary structures with pseudoknots*, Nucleic Acids Research (2011) Link to article

Feedback

To report errors in accessing the webserver please contact: anne.capdepon@cea.fr .

HDV RNA
S=93%
PPV=84%



genus = 2

```
UGGCCGGCAUGGUCCCAGCCUCCUCGCUGGCGCCGGCUGGGCAACAUUCCGAGGGGACCGUCCCCUCGGUAAUGGCGAAUGGGACCCA
UGGCCGGCAUGGUCCCAGCCUCCUCGCUGGCGCCGGCUGGGCAACAUUCCGAGGGGACCGUCCCCUCGGUAAUGGCGAAUGGGACCCA
((.(((((.((((.(((((.....)))))).)))))))))).)).(((((((((((((((.(((((((((.....(((...[[[[.
[[[[[.(((((([[[.
[[[[[[[[.....)))))))))]]]]]]].].]]].)))))))))).))))))))))).))))....]]]]].]]]]
```

http://ipht.cea.fr/rna/mcgenus.php

What about real knots?

- In polymer, probability of unknot:

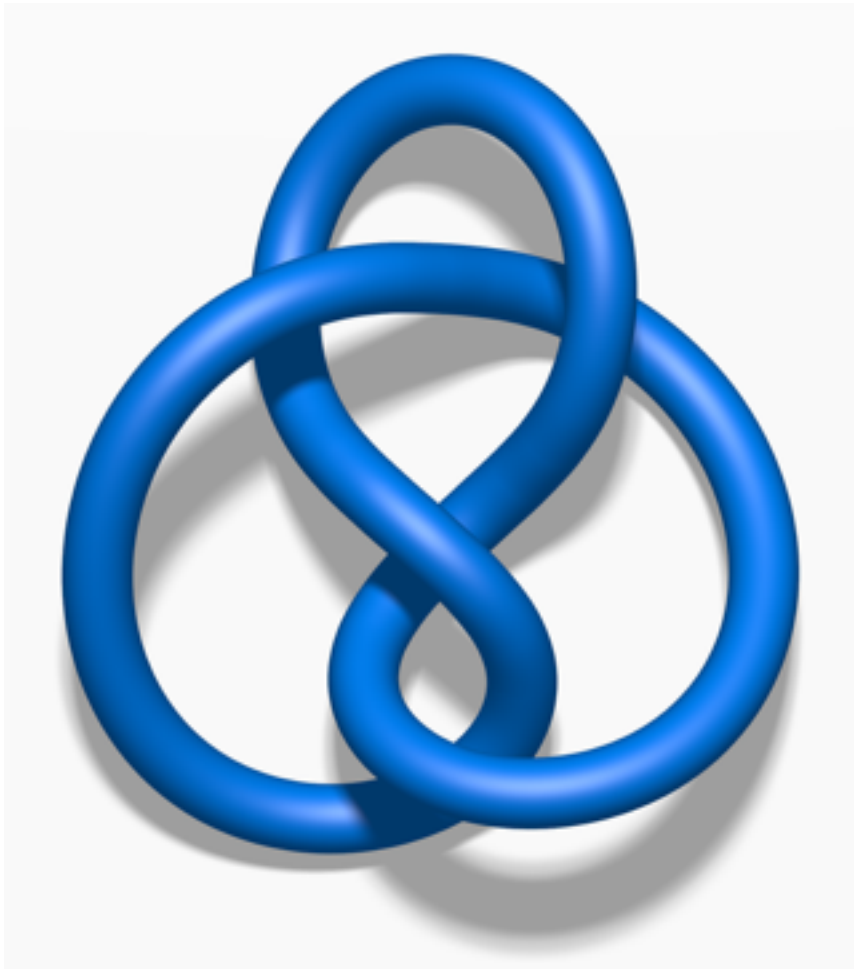
$$P_0(n) = e^{-(n-n_0)/n_c}$$

- Very frequent in ds DNA (viral) and very complex (up to 20 minimal crossings)
- Around 2% of all the PDB proteins are knotted (mostly trefoil but one 6-knot)
- What about RNA???

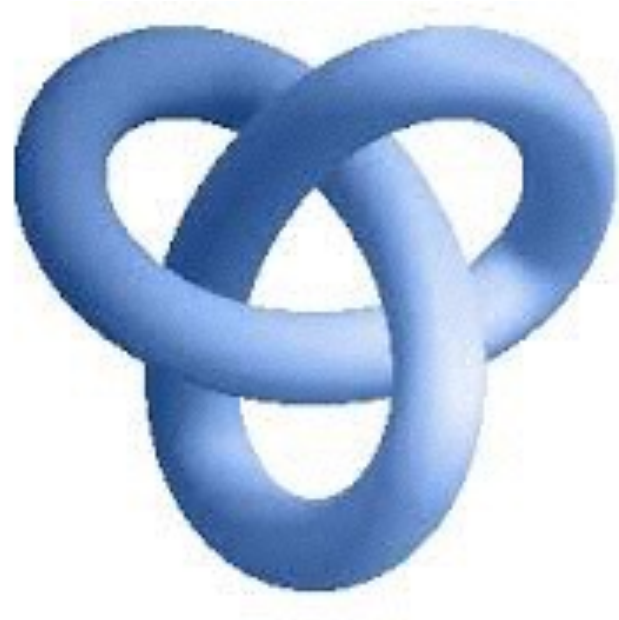
- Look into PDB: 1041 RNA alone, 1801 hybridized
- In total, 6219 distinct RNA chains
- Each chain circularized using the minimally invasive scheme
- Compute Alexander polynomial and Dowker code

- Only three knotted structures!
  - a 16 crossing prime knot in 3JYX5 (comprising 3170 nucleotides)
  - a 4_1 prime knot in 2GYA0 (comprising 2740 nucleotides)
  - a figure of 8 knot and three trefoil knots in 1C2W:B

all solved by cryo-em

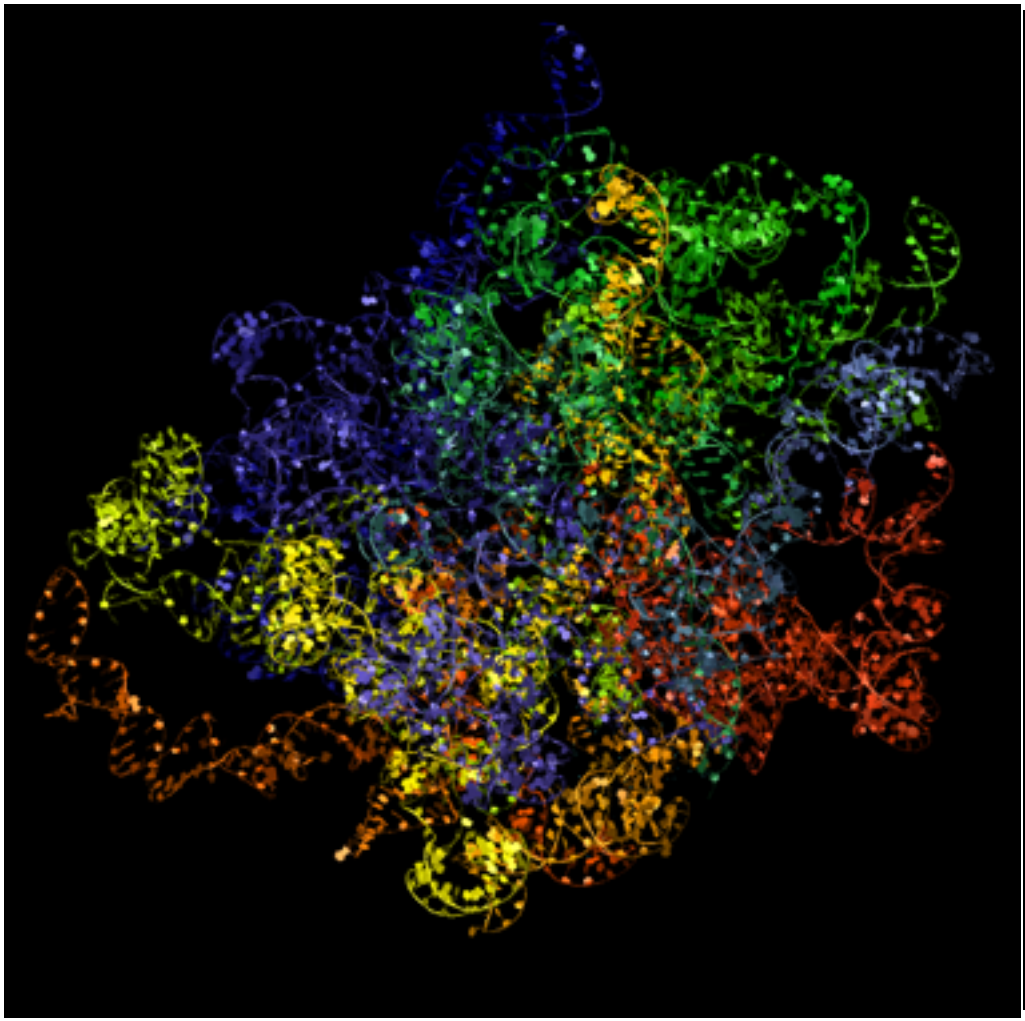


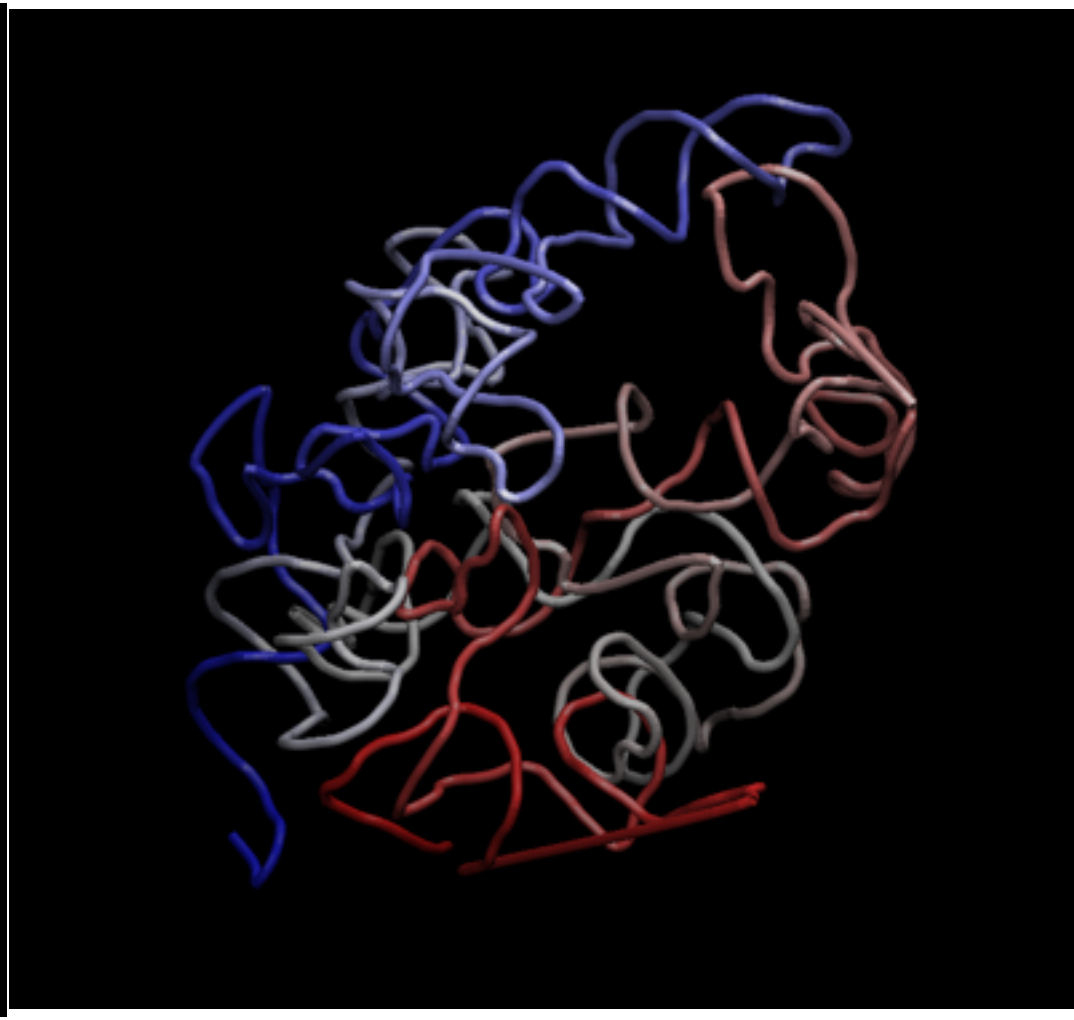
The 4_1 knot



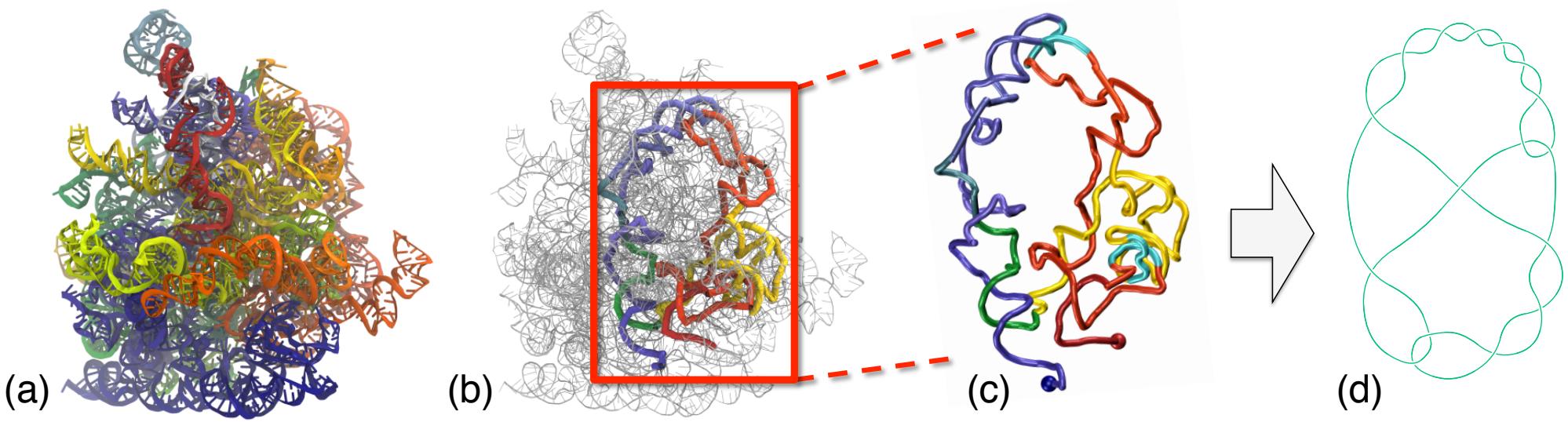
The Trefoil knot

3JYX5



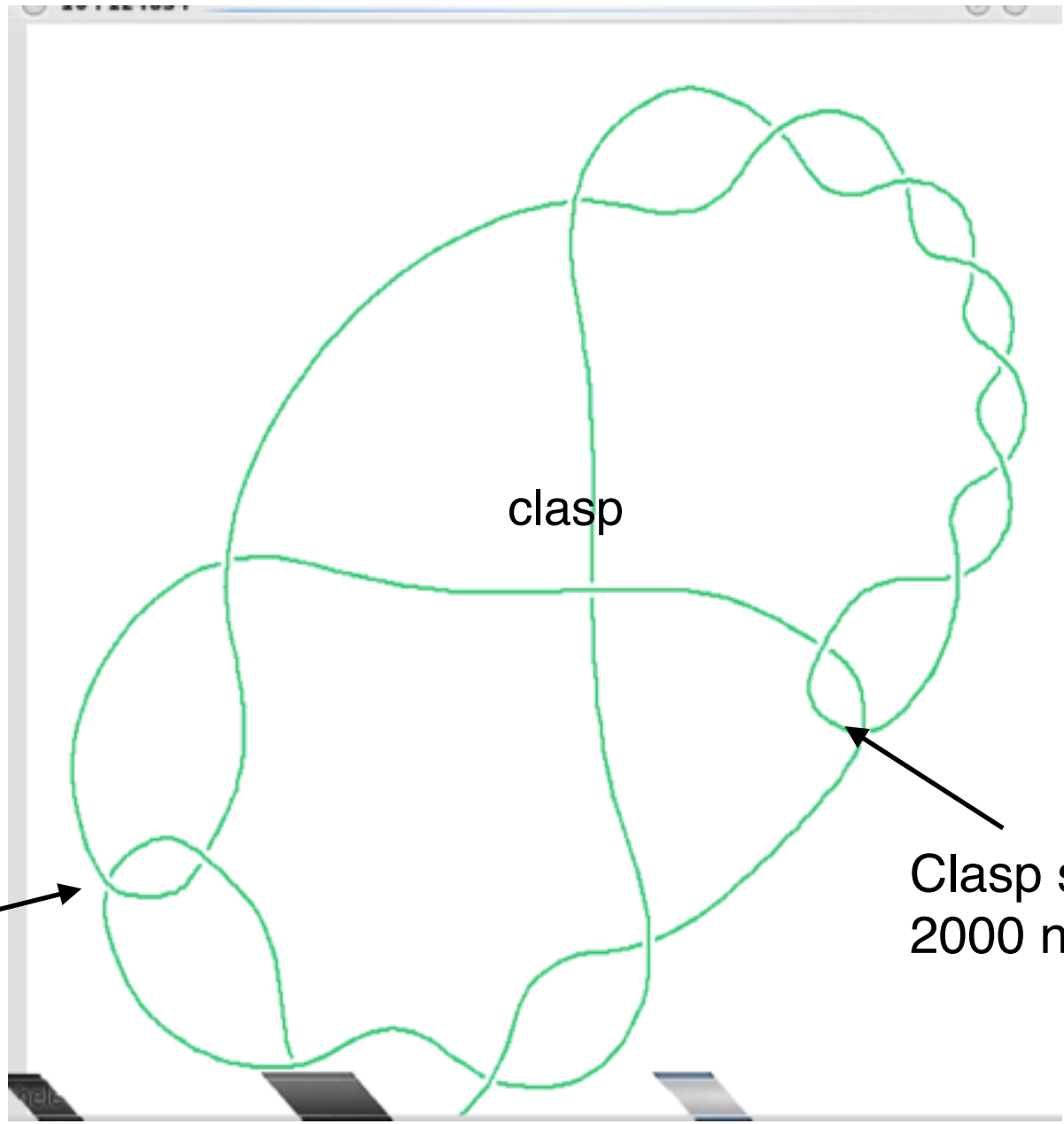


Genus=5

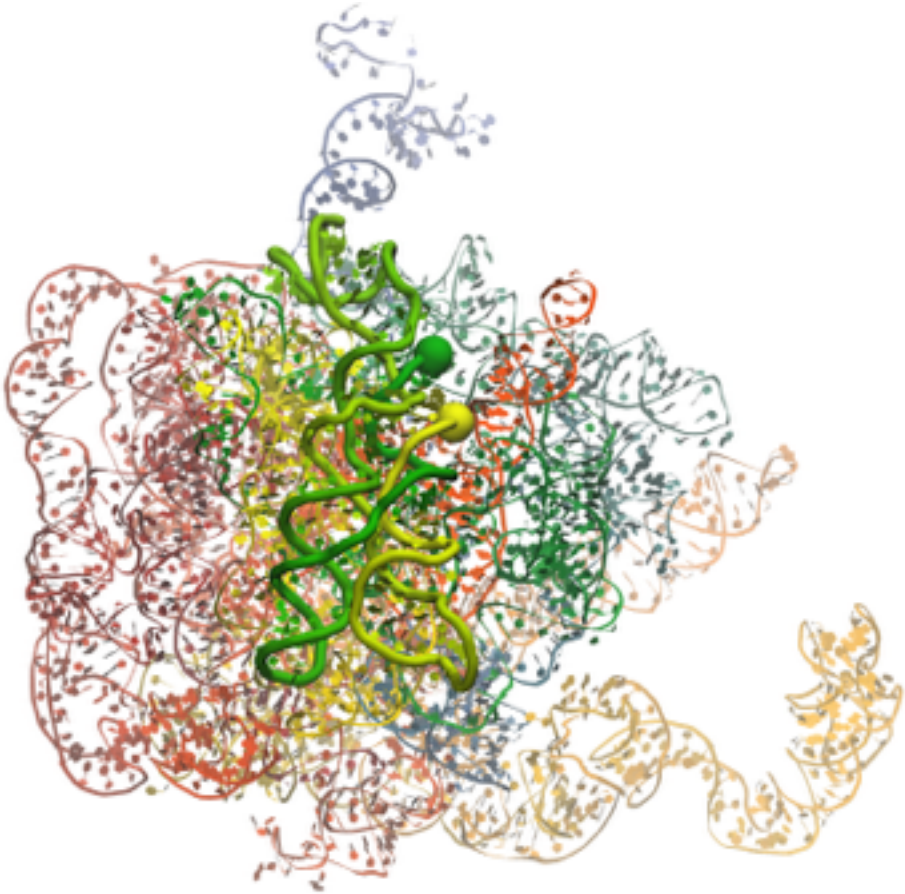


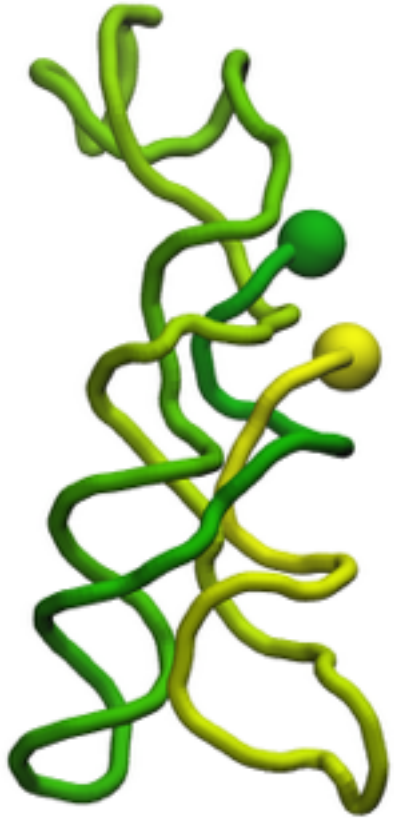
Knotted 26S ribosomal RNA structure from PDB entry 3JYX:5.

3JYX5



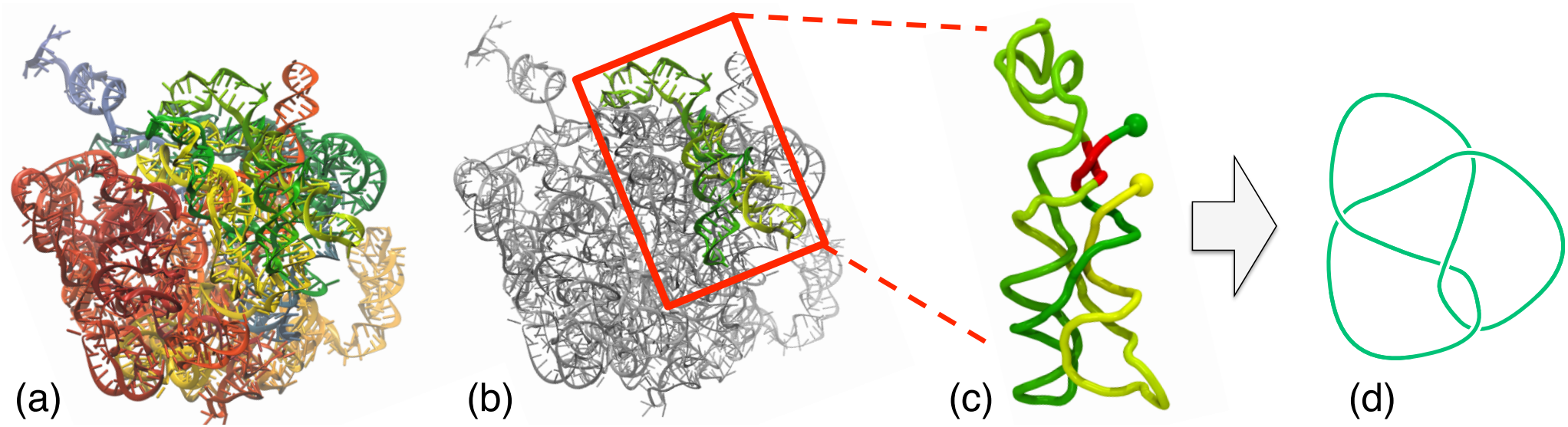
2GYA0



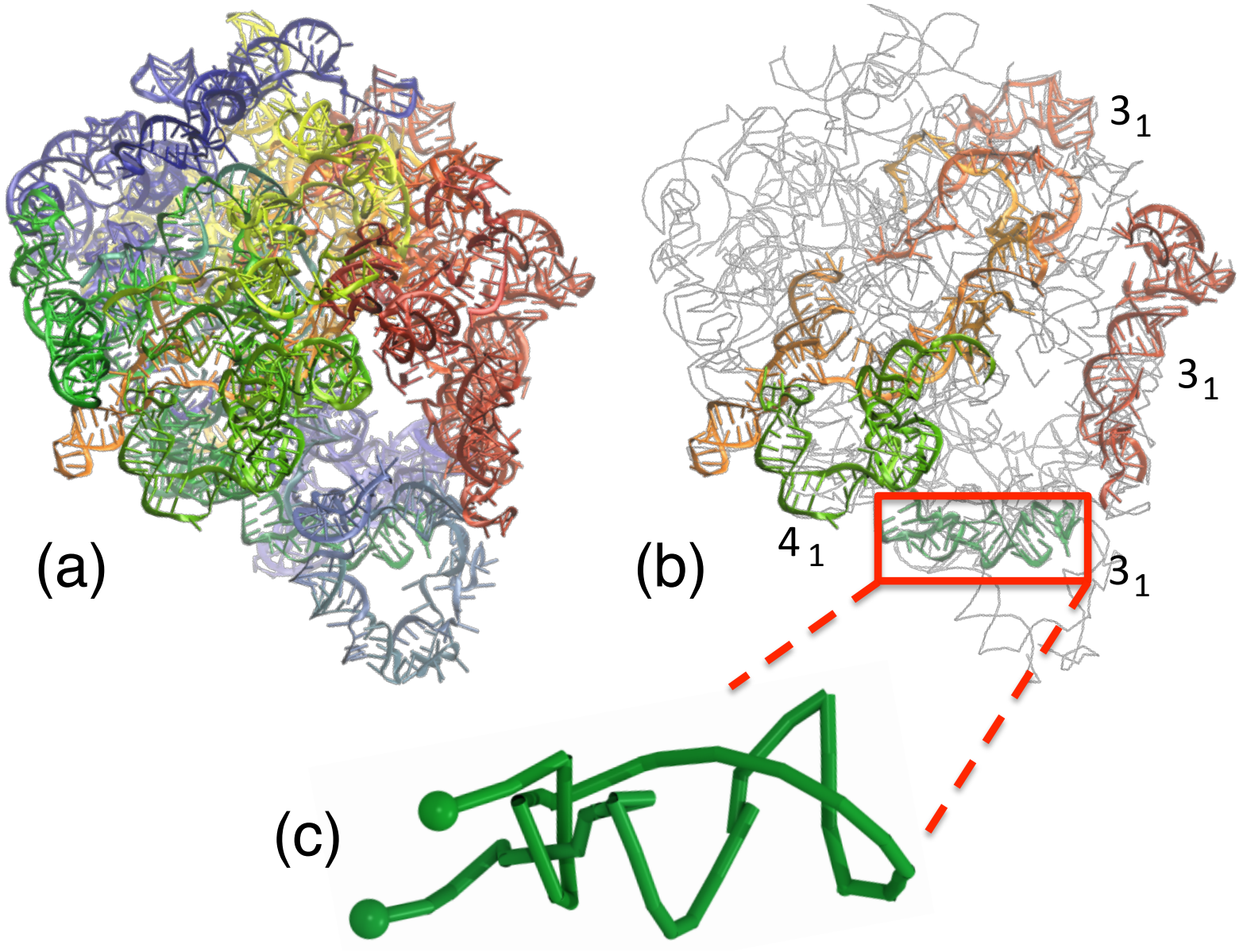


Achiral twist knot
150 nt knot

Genus=6



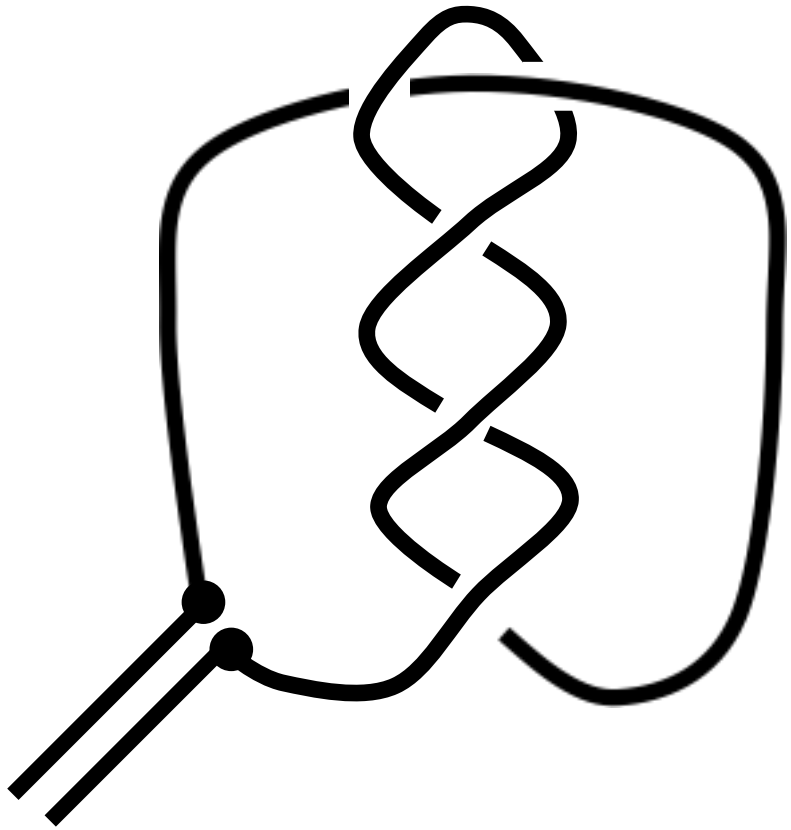
Knotted 23S ribosomal RNA structure from PDB entry 2GYA:0.



Knotted 23S ribosomal RNA structure from PDB entry 1C2W:B.

- All structures from cryo-em
- There is probably an error in the structure of 3JYX5
- 2GYA0 and 1C2W may have a genuine knot, but again could be an artefact of structure resolution since very close homologs have no knots
- Conclusion: knots are very rare in RNA, and possibly non-existent!

Design of RNA knots



Design of RNA twist knots

Conclusion

- One needs a refined energy model to improve predictions
- Need to include steric constraints at an early stage in the algorithm
- Are there knots in RNA?

Probably (k)not!