A topological classification of pseudoknots in RNA

Henri Orland (SPhT, Saclay)
Collaboration with

- A. Zee (KITP, UCSB)
- and
- G. Vernizzi (Northwestern University)
- M. Bon (Saclay)
Outline

- Secondary structures
- Matrix field theory for RNA
- Topological classification of RNA
- Exact enumeration of RNA structures
- Monte Carlo approach
Pseudoknots in RNA

- In addition to planar secondary structures, there are “pseudoknots” which constrain the 3d structure
Big simplifications

- **saturation** of Crick–Watson pairing (or stacking)
- number of bases in **pseudoknot**
- $\ll$ number of bases in **planar secondary structure**.
\[ Z = \sum Q_0 \]

sterically allowed configurations

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

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

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

- any index appears once and only once (\textit{saturation})

Base pair energy

Chain rigidity
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 $L$

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

- **due to Wick theorem**

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

\[ V_{ij} V_{kl} + V_{ik} V_{jl} + V_{il} V_{jk} = \frac{1}{N} \int \prod_{i=1}^{L} d\phi_i e^{-\frac{1}{2} \sum_i \sum_j \phi_i V_{ij}^{-1} \phi_j \phi_i \phi_j \phi_k \phi_l} \]

- However, this form gives same weight to all pairings. No penalty for Pseudoknots.
- Experimentally, few pseudoknots.
• We look for a parameter $N$ such that

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

• Corrections in $\frac{1}{N} \equiv \text{Pseudoknots}$

• **TOPOLOGY=MATRIX FIELD THEORY**

t’Hooft 1972
Matrix field representation of RNA folding

- We thus generalize the Wick theorem
  - scalar fields
    \[ Z(1, L) = \frac{1}{N} \int d\phi_i e^{-\frac{1}{2} \sum_{i,j} \phi_i V_{ij}^{-1} \phi_j} \prod_{i=1}^{L} (1 + \phi_i) \]
  - matrix fields
    \[ Z(1, L) = \frac{1}{A(L)} \int d\varphi_k e^{-\frac{N}{2} \sum_{i,j} (V^{-1})_{ij} \text{tr}(\varphi_i \varphi_j)} \frac{1}{N} \text{tr} \prod_{l=1}^{L} (1 + \varphi_l) \]
• at each site, there is a symmetric $N \times N$ matrix over which one integrates:

$$\varphi_{ab}(i) \text{ with } i = 1, \ldots, L$$

• in standard Feynman graphs, use single lines

• for matrix theory, use double lines, carrying indices $(a,b)$
Double line graphs

- In our problem, if we use matrix fields

\[ \phi_{ab}(x) : \text{NxN matrix} \]

- Diagram rules:

- 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: $1$
- Order $1/N^2$
• By looking at a few diagrams: planar diagrams are of order 1 in $1/N$, pseudo-knots are of higher order.

• Matrix field theory seems to do what we want: discriminate PK.
the MATRIX secretly rules
• It can be shown that in matrix theory, each RNA graph is weighted by a factor

\[
\frac{1}{N^{2g}}
\]

where \( g \) is the topological genus of the graph.

• In fact, the matrix field partition function is equal to

\[
Z = \sum_{\text{all pairings}} \frac{1}{N^{2g}\text{(pairing)}} e^{-\beta E\text{(pairing)}}
\]
What is the genus?

• Secondary structures without PK can be drawn **without crossings** on a sphere.

• An RNA graph has genus $g$ if it can be drawn **without crossings** on a closed surface with at least $g$ handles.

• For an RNA graph, the genus is

\[ g = \frac{\text{Pairings} - \text{Loops}}{2} \]
Topological classification of RNA folds

- An RNA fold can be characterized by its topological genus
- Number of handles of embedding surface
Figure 3: First few terms of the topological expansion of closed oriented surfaces: the term $g = 0$ is ... Note that the circle of the RNA-backbone (in green) topologically corresponds to a hole (or puncture) on the surface.
Genus 0: the Sphere
Genus 1: the Torus
Genus 2: the Bi-torus
Genus 3
Large N expansion

- One can show that the large N limit of the matrix integral reproduces exactly the recursion for the secondary structures (genus 0)
- One can obtain a recursion equation for graphs of genus 0 and 1.

Algorithmic complexity scales like $L^6$

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

ABCABC
<table>
<thead>
<tr>
<th>EMBL Accession number</th>
<th>Description</th>
<th>Start</th>
<th>End</th>
<th>Bits Score</th>
<th>Get Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB032408.1</td>
<td>Shewanella violacea genes for SecY, ribosomal protein S13, ribosomal ...</td>
<td>554</td>
<td>653</td>
<td>69.9400</td>
<td></td>
</tr>
<tr>
<td>AE004325.1</td>
<td>Vibrio cholerae O1 biovar eltor str. N1961 chromosome I, section 233...</td>
<td>10782</td>
<td>10672</td>
<td>80.2700</td>
<td></td>
</tr>
<tr>
<td>AE005556.1</td>
<td>Escherichia coli O157:H7 EDL933 genome, contig 3 of 3, section 175 of...</td>
<td>8956</td>
<td>8845</td>
<td>102.2000</td>
<td></td>
</tr>
<tr>
<td>AE006177.1</td>
<td>Pasteurella multocida subsp. multocida str. Pm70 section 144 of 204 o...</td>
<td>3108</td>
<td>2993</td>
<td>88.4700</td>
<td></td>
</tr>
<tr>
<td>AE008857.1</td>
<td>Salmonella typhimurium LT2, section 161 of 220 of the complete genome...</td>
<td>15653</td>
<td>15542</td>
<td>100.2500</td>
<td></td>
</tr>
<tr>
<td>AE014003.1</td>
<td>Yersinia pestis KIM section 403 of 415 of the complete genome.</td>
<td>5850</td>
<td>5961</td>
<td>97.1300</td>
<td></td>
</tr>
<tr>
<td>AE015343.1</td>
<td>Shigella flexneri 2a str. 301 section 306 of 412 of the complete geno...</td>
<td>6155</td>
<td>6044</td>
<td>102.2000</td>
<td></td>
</tr>
<tr>
<td>AE015474.1</td>
<td>Shewanella oneidensis MR-1 section 23 of 457 of the complete genome.</td>
<td>23</td>
<td>120</td>
<td>68.2200</td>
<td></td>
</tr>
<tr>
<td>AE016767.1</td>
<td>Escherichia coli CFT073 section 13 of 18 of the complete genome.</td>
<td>243541</td>
<td>243430</td>
<td>102.2000</td>
<td></td>
</tr>
<tr>
<td>AE016799.1</td>
<td>Vibrio vulnificus CMCP6 chromosome I section 3 of 11 of the complete ...</td>
<td>152117</td>
<td>152009</td>
<td>77.1800</td>
<td></td>
</tr>
<tr>
<td>AE016848.1</td>
<td>Salmonella enterica subsp. enterica serovar Typhi Ty2, section 15 of ...</td>
<td>21171</td>
<td>21282</td>
<td>100.2500</td>
<td></td>
</tr>
<tr>
<td>AE016992.1</td>
<td>Shigella flexneri 2a str. 2457T section 15 of 16 of the complete geno...</td>
<td>262117</td>
<td>262228</td>
<td>102.2000</td>
<td></td>
</tr>
<tr>
<td>AE017127.1</td>
<td>Yersinia pestis biovar Medeavals str. 91001 section 1 of 16 of the c...</td>
<td>231109</td>
<td>231220</td>
<td>97.1300</td>
<td></td>
</tr>
<tr>
<td>AE017156.1</td>
<td>Haemophilus ducreyi strain 35000HP section 6 of 6 of the complete gen...</td>
<td>132595</td>
<td>132483</td>
<td>79.8900</td>
<td></td>
</tr>
<tr>
<td>AJ414141.1</td>
<td>Yersinia pestis strain CO92 complete genome; segment 1/20</td>
<td>232397</td>
<td>232508</td>
<td>97.1300</td>
<td></td>
</tr>
<tr>
<td>AL627282.1</td>
<td>Salmonella enterica serovar Typhi (Salmonella typhi) strain CT18, com...</td>
<td>5691</td>
<td>5802</td>
<td>100.2500</td>
<td></td>
</tr>
<tr>
<td>AP002564.1</td>
<td>Escherichia coli O157:H7 DNA, complete genome, section 15/20.</td>
<td>266263</td>
<td>266152</td>
<td>102.2000</td>
<td></td>
</tr>
<tr>
<td>AP005073.1</td>
<td>Vibrio parahaemolyticus DNA, chromosome 1, complete sequence, 1/11.</td>
<td>278188</td>
<td>278299</td>
<td>72.3400</td>
<td></td>
</tr>
<tr>
<td>AP005331.1</td>
<td>Vibrio vulnificus Y0016 DNA, chromosome I, complete genome, section 2...</td>
<td>149281</td>
<td>149389</td>
<td>77.1800</td>
<td></td>
</tr>
<tr>
<td>BX571874.1</td>
<td>Pseudoalteromonas alutagens subsp. lacrimaria strain TT01 complete genome; segment 2</td>
<td>272732</td>
<td>272618</td>
<td>82.0900</td>
<td></td>
</tr>
<tr>
<td>BX590851.1</td>
<td>Erwinia carotovora subsp. atroseptica strain 1043, complete genome</td>
<td>4490616</td>
<td>4490508</td>
<td>86.6200</td>
<td></td>
</tr>
<tr>
<td>CR378663.1</td>
<td>Photobacterium profundum S59; segment 1/12</td>
<td>343376</td>
<td>343482</td>
<td>61.9700</td>
<td></td>
</tr>
<tr>
<td>XD2543.1</td>
<td>E. coli alpha ribosomal protein operon for ribosomal proteins S13, S1...</td>
<td>141</td>
<td>252</td>
<td>102.2000</td>
<td></td>
</tr>
<tr>
<td>M12432.1</td>
<td>E. coli alpha operon ribosomal protein S13 (rpsM) gene, 5' end and pro...</td>
<td>572</td>
<td>683</td>
<td>102.2000</td>
<td></td>
</tr>
<tr>
<td>U18097.1</td>
<td>Escherichia coli K-12 chromosomal region from 67.4 to 76.0 minutes.</td>
<td>223298</td>
<td>223187</td>
<td>102.2000</td>
<td></td>
</tr>
<tr>
<td>U32762.1</td>
<td>Haemophilus influenzae Rd KW20 section 77 of 163 of the complete geno...</td>
<td>6695</td>
<td>6810</td>
<td>91.4500</td>
<td></td>
</tr>
<tr>
<td>U00096.2</td>
<td>Escherichia coli K-12 MG1655 complete genome.</td>
<td>3440572</td>
<td>3440461</td>
<td>102.2000</td>
<td></td>
</tr>
</tbody>
</table>

H. Orland, SPhT, Saclay

27 potential ABCABC
Statistical study

• Look in database and calculate genii of pseudoknots

• PseudoBase: around 245 primitive pseudoknots

• 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 are KHP of genus 1
Number of RNA as a function of genus
Figure 10: A typical tRNA (PDB ID 1evv [34]). It has the genus 1 of a kissing hairpin pseudoknot.
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

\[
Z_N(L) = \frac{1}{A} \int d\phi \ e^{-N \frac{\text{Tr} \phi^2}{2v}} \frac{1}{N} \text{Tr} (1 + \phi)^L
\]

- with only one \( N \times N \) matrix \( \phi \)
• 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

\[ \langle g \rangle_L \approx 0.25L \]

• 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
  \[ <g> \approx 0.13L \]
- 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

• With current Turner energies, we are not able to find the correct pseudoknots in RNA
  • transfer RNAs (g=1)
  • Hepatitis delta virus ribozyme (g=2)
• We find pseudoknots with lower free-energies than the native ones!
• We are working on a new parametrization of free energies in RNA able to discriminate real PK.
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: the Monte Carlo calculation with chemical potential for the genus.