Glassiness in RNA secondary structures

Ralf Bundschuh
The Ohio State University

Collaborators: Terence Hwa, University of California at San Diego
Tsunglin Liu, The Ohio State University
Ulrich Gerland, TU München

Outline:

• Introduction to RNA
• Uniform sequences: the molten phase
• Disorder: glass phase and glass transition ?
• Bias: the native phase
• Finite size effects
• Quantitative modeling
• Conclusions

supported by DAAD, Beckman foundation, and NSF
**Introduction**

- RNA is **heteropolymer** of four different bases G, C, A, and U

![RNA bases and structure](image)

- **Primary structure**: Sequence, e.g.,
  
  GCGGAUUUAGCUCAGUUGGAGAGCGCCAGACUGAAAUCCUGUCGUGUUCGAGAUCACUCGACCA

- **Strongest interaction**: Watson-Crick **base pairing** (G–C and A–U)
  
  → secondary structure

- **Spatial arrangement**
  
  → tertiary structure

(looks locally like DNA double helix)
Introduction II

• Secondary structure: Set of base pairs formed

• Pseudo-knots neglected

• Diagramatic representation

• Assign energy $E[S]$ to each structure $S \rightarrow$ partition function

$$Z = \sum_{\{S\}} \exp(-E[S]/T)$$

• Partition function generated exactly by Hartree equation

$$i \quad j = \quad i \quad j-1 \quad j + \sum_k \quad i \quad k \quad j-1 \quad j$$

→ Electron in disordered medium, meanders

• Handle for analytical treatment

• $O(N^3)$ algorithm for exact partition function (McCaskill, Biopolymers 29, 1990.)
Introduction III

- Two important parameters for generic properties:
  - Temperature
  - Bias for native structure

- Use long hairpin as native (designed) structure

- Create sequences by
  - randomly choosing first half of the sequence
  - assigning exact complement as second half
  - changing bias by mutations with probability $p$

- Expected phase diagram:
• Are there really three distinct compact phases?
• Characterize all phases numerically by scaling laws
• Choose random sequences and calculate for each of them their typical size $\langle h \rangle$ numerically
• Different behavior in all three compact phases:
De Gennes (1968) proposed: start with uniform sequences
AUAAUAUAUAU... or GCGCGCGCGC...

Uniform attraction between any two elements of the polymer → only one effective interaction parameter $\varepsilon_0$

$\varepsilon_0$ contains binding energy and entropic terms relative to unbound RNA

Most monomers engaged in base pairs

Main effect in molten phase: branching entropy

Hartree equation becomes

$$\hat{Z}(\mu)^{-1} = (e^\mu - 1) - \hat{\Pi}(\mu) \quad \hat{\Pi}(\mu) = e^{-\varepsilon_0/T} \hat{Z}(\mu)$$

in Laplace domain → $Z(N) \sim N^{-\theta} e^{\mu_0 N}$ with $\theta = \frac{3}{2}$

Molten Phase II

- **mountain representation** \((N = 18)\)

![Mountain representation diagram](image)

- **one to one correspondence**: RNA secondary structures \(\leftrightarrow\) mountains

- all bases attract **equally strong**
  - counting structures \(\leftrightarrow\) counting mountains
  - free random walk in presence of a hard wall

\[
\rightarrow Z(N) \sim N^{-3/2} e^{\mu_0 N}
\]
\[
\rightarrow \langle h \rangle \sim N^{1/2}
\]
\[
\rightarrow R_g \sim N^{1/4}
\]

\(\rightarrow\) branched polymer
• Is molten phase stable? → perturbative approach

• Interaction energy between base $i$ and base $j$

$$\varepsilon_{ij} = \varepsilon_0 + \Delta \varepsilon_{ij}$$

• Assume $\Delta \varepsilon_{ij}$ independent Gaussian variables

$$\overline{\Delta \varepsilon_{ij}} = 0 \quad \overline{\Delta \varepsilon_{ij} \Delta \varepsilon_{kl}} = \Delta \varepsilon \, \delta_{ik} \delta_{jl}$$

• First term in free energy expansion in powers of $\Delta \varepsilon$: two-replica system $\overline{Z^2}$

→ two replicas which gain energy $\Delta \varepsilon$ for every common bond (●····●)
Glass transition II

- Two replicas which gain energy $\Delta \varepsilon$ for every common bond (---)

Order configurations of 2 replica system by configurations of common bonds

Common bonds (---) form RNA structure themselves

represents sum over all possible choices of non-common bonds in the two replicas

1 replica $\longrightarrow \ell^{-3/2}$ $\Rightarrow$ 2 replicas $\longrightarrow (\ell^{-3/2})^2 = \ell^{-6/2}$

effective picture: single RNA with “6-dimensional” loop entropies
Glass transition III

- exactly solvable
- phase transition at finite $\Delta \varepsilon_c$
  - $\Delta \varepsilon < \Delta \varepsilon_c$: 2 replicas fluctuate independently (molten)
  - $\Delta \varepsilon > \Delta \varepsilon_c$: 2 replicas have same contacts (glass)
  - specific heat exponent $\alpha = 1$
    $\rightarrow$ marginally first order transition

- fraction of common contacts agrees well with Monte Carlo simulations

$\Rightarrow$ molten phase perturbatively stable
• Is glass phase stable?

• Study pinching excitations

• Assume molten phase is stable for all temperatures

$$Z(T, N) = A(T) N^{-3/2} \exp[-f_0(T)N]$$

Tang and Chaté, PRL 86 (2001)

• On the one hand:

$$\Delta F(N) = -2T \log(N/2)^{-\frac{3}{2}} + T \log N^{-\frac{3}{2}} \approx \frac{3}{2} T \log N$$
Glass Transition V

- On the other hand: find piece of length $\log N / \log 2$ in first half *exactly complementary* to piece in second half

  ⇒ All pairs in this piece contribute pairing energy $\epsilon_P$ in unpinched configuration

- Estimate for pinching free energy:

  $$\Delta F(N) \geq [\epsilon_P + 2f_0(T)] \log N / \log 2$$

- Combine two results:

  $$\frac{3}{2} T \geq [\epsilon_P + 2f_0(T)] / \log 2$$

- $\epsilon_P + 2f_0(T \to 0) > 0$ (for four or more base alphabet)

  ⇒ contradiction for $T < T_*$

  ⇒ RNA *cannot* be in molten phase for $T < T_*$

  ⇒ different (glass) phase at low temperatures
Molten-Native Transition I

- How does native structure appear?
- Start from molten phase
- Add bias towards native structure (hairpin)
- Use simplified model in spirit of molten phase description
- Possible Watson-Crick pairs:

- only two different interaction energies
- strong interaction $\varepsilon_0 + U_0$ for native base pairs
- weak interaction $\varepsilon_0$ for all other base pairs
- $U_0$ is an effective measure of the bias

Molten-Native Transition II

- Model can be exactly solved by Laplace transform
- Phase transition between molten and native phase at finite critical bias $U_c$ or at critical temperature $T_c$
- Phase transition is second order with finite jump in specific heat ($\alpha = 0$) but large finite size effects possible
- Calculate fraction of native contacts $Q$
- Exhibits scaling form and scaling function

$$Q \sim N^{-1/2} g \left( \frac{T - T_c}{T_c} N^{1/2} \right)$$

($\nu = 2$) where

$$g(y) \approx \begin{cases} 
-y^1 & y \ll -1 \\
1 & -1 \ll y \ll 1 \\
y^{-1} & y \gg 1 
\end{cases}$$

- Can be numerically verified to apply to randomly chosen RNA sequences
Finite size effects I

- Theory in thermodynamic limit \( N \to \infty \)
- Natural RNA \( N \approx 100 - 10000 \)
- Numerical calculations: single seq. \( N \approx 10000 \), ensemble average \( N \approx 1000 \)
- Is asymptotic theory applicable?
- Back to molten phase, uniform interaction \( \varepsilon_0 \), Boltzman factor \( q = e^{-\varepsilon_0/kT} \)
- Partition function in Laplace domain:
  \[
  \hat{Z}(\mu)^{-1} = (e^\mu - 1) - q \hat{Z}(\mu)
  \]

- Can extract not only leading behavior but also
  \[
  Z(N) = A(q) N^{-\theta} e^{\mu_0(q) N} \left[ 1 + \frac{N_0(q)}{N} + O \left( \frac{1}{N^2} \right) \right]
  \]
  with
  \[
  \theta = \frac{3}{2}, \quad A(q) = \left[ \frac{(1 + 2\sqrt{q})}{4\pi q^{3/2}} \right]^{1/2}, \quad \mu_0(q) = \ln(1+2\sqrt{q}), \quad N_0(q) = \frac{3(1 + 4\sqrt{q})}{16\sqrt{q}}
  \]
- \( N_0(q) \) is cross-over length
Finite size effects II

- Cross-over length is nonuniversal quantity

⇒ Need more quantitative model!
⇒ Include loop entropies
- Partition function still given by quadratic equation in Laplace space

⇒ cross-over length $N_0$ can be calculated, large $s$ expansion:

$$N_0 \approx \frac{3s^{\frac{3}{4}}}{8\sqrt{hm}}$$

- Use measured parameters for repeated sequences

<table>
<thead>
<tr>
<th>Sequence</th>
<th>$N_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(AU)$_n$</td>
<td>8000</td>
</tr>
<tr>
<td>(GC)$_n$</td>
<td>16000</td>
</tr>
<tr>
<td>(CAG)$_n$</td>
<td>7</td>
</tr>
</tbody>
</table>

- Finite size effects strongly sequence dependent
- Finite size effects can be very large!
RNA secondary structure difficult to determine experimentally since RNA is very floppy ⇒ no two copies of an RNA have the same shape

Alternative: single-molecule experiments

- Attach ends of RNA molecule to two beads
- Keep beads at fixed distance $R$ with optical tweezers
- Measure force $f$ on beads as function of distance $R$
- P5ab hairpin of tetrahymena thermophila group I intron

Liphardt, Onoa, Smith, Tinoco, and Bustamante, Science, 2001

Method to learn something about secondary structure?
Quantitative modeling II

- Experiments hard ⇒ quantitative modeling

- Two ingredients: secondary structure and polymer physics of backbone

- Secondary structure:
  - Need partition function $Q(m)$ of RNA molecule given $m$ “exterior” bases
  - Can be calculated by modifying “Vienna package”
  - Uses detailed free energy rules

- Backbone physics
  - Elastic freely jointed chain
  - Persistence length $1.9nm$/base distance $0.7nm$
  - Partition function at distance $R$ and length $m$: $W(R;m)$

- Total partition function:

$$Z(R) = \sum_{m=0}^{N} Q(m)W(R;m)$$
Apply to P5ab hairpin of tetrahymena thermophila group I intron

- **Quantitative prediction**
- **Disadvantages** compared to experiment:
  - No tertiary structure or pseudo-knots
  - Parameters not exactly known
- **Advantages** over experiment:
  - Can be rapidly applied to arbitrary sequence
  - Intermediate structures can be investigated
- Offered as interactive web server

http://bioserv.mps.ohio-state.edu/rna
Quantitative modeling IV

- Apply to full *tetrahymena thermophila* group I intron
- Group I intron contains **pseudo-knot**!
- Quantitative modeling ignores pseudo-knot
  \[ \Rightarrow \text{known inactive conformation} \]
- No sign of secondary structure!

- **Consistent** with experiments on single-stranded DNA

  Maier, Strick, Croquette, and Bensimon, Single Molecules, 2000
• What’s happening?

• Look at intermediate structure (here $R = 100 nm$)

• “Socks on the clothes line”

  • Compensation effect:
    – Extension $R$ is increased
    – One of the “socks” disappears
    – The other “socks” take up the slag

  ⇒ No rapid change in force as sock disappears
  ⇒ smooth force-extension curve
• What can be done?
• Add spring

Spring constant $\lambda$
• Vary $R_t$
• Measure extension $\langle R \rangle$ and fluctuations

$$\delta R \equiv \sqrt{\langle (R - \langle R \rangle)^2 \rangle}$$
• Fluctuations show structure

Peaks in fluctuations correlated with disappearance of secondary structure elements
• Alternative experimental setup: pulling through a nano-pore

"cis"  "trans"

n nucleotides

• Kinetic model: diffusion in a time dependent one-dimensional energy landscape ⇒ Monte Carlo simulation

• Signature of every structural element
• RNA shows a large variety of interesting behavior.

• RNA secondary structure formation is tractable analytically and numerically by methods of statistical mechanics.

• Quantitative description of single-molecule experiments possible.

• Single-molecule experiments reveal structure information through measurements of fluctuations or with the help of a nano-pore.
Conclusions and Outlook II

Future work:

- Understand glass phase properties analytically
- glass-native transition
- more realistic RNA models, self-avoidance
- interactions between several molecules
- Comparison of quantitative predictions with experiments on medium-sized RNA
- kinetics
- pseudo-knots
- tertiary structure
- biological applications (RNA finding, Huntington’s disease)
Biological function of RNA

• Biological functions:

  • **Structure** (→ proteins)
    – Ribosomal RNA
    – Transfer RNA

  • **Information** (→ DNA)
    – Messenger RNA
    – single-stranded DNA
      * T instead of U
      * more rigid backbone

• **Interplay** of structure and information
  – Splicing
  – Ribozymes
  – RNA world (origin of life)
Molten Phase in Natural Molecules I

- Application to real sequences
- Use experimentally determined parameters from RNA secondary structure prediction which take all energetic details into account

Uniform sequences $\text{AUUAUUAUUAU}\ldots$ and $\text{GC}GCG\text{CGC}-\text{GC}\ldots$ need very long sequences ($\geq 8000$ bases)
  - Hairpin loops must contain at least 3 bases
  - Loss in binding energy large in hairpin loops

Hofacker et al., Monatshefte f. Chemie 125, 1994
Molten Phase in Natural Molecules II

- Naturally occurring in human DNA: \((CAG)_n\) with large \(n\)
- Connected with Huntington’s disease
- Hereditary neurodegenerative disease
  - \(n < 35\) normal
  - \(n > 35\) Huntington’s disease
- If \(n > 35\), \(n\) usually very large
- CAG codes for Glutamine \(\rightarrow\) repeats appear in protein
- Single-stranded DNA can undergo self-binding during replication
- Biologist’s model: only minimal free energy structure competes with single-stranded configuration
• (CAG)$_n$ can be in molten phase

• Crossover length 7 bases

⇒ Molten phase relevant

• Kinetics possibly important

• (single molecule ?) experiments necessary
Solution of Gō model

- Partition function: order arbitrary structure by number of **native**
  contacts

\[
Z(N, U_0) = \sum \cdots
\]

- \(\tilde{W}(\ell)\) = sum over all ways to place non-native bonds

- \(W\) similar to molten phase partition function \(\longrightarrow\) expect \(W(\ell) \sim \ell^{-3/2}\)

- Relation between bubble \((\tilde{W})\) and full \((Z)\) partition functions

\[
\hat{Z}(\mu; U_0) = \hat{W}(\mu) + \hat{W}(\mu) e^{-(\varepsilon_0 + U_0)/T} \hat{W}(\mu) \\
+ \hat{W}(\mu) e^{-(\varepsilon_0 + U_0)/T} \hat{W}(\mu) e^{-(\varepsilon_0 + U_0)/T} \hat{W}(\mu) + \ldots
\]

(Laplace domain)

\[
\longrightarrow \hat{Z}(\mu; U_0)^{-1} = \hat{W}(\mu)^{-1} - e^{-(\varepsilon_0 + U_0)/T}
\]

- Exact expression for \(\hat{Z}(\mu; U_0)\)

- Partition function relation \(\hat{Z}(\mu; U_0)^{-1} = \hat{W}(\mu)^{-1} - e^{-\frac{\varepsilon_0 + U_0}{T}}\)

- Boundary condition \(Z(N, U_0 = 0) = Z(2N)\) gives \(\hat{W}\)
Molten-Native Transition III

- Free energy as a function of the number of total contacts $K$ and the number of native contacts $Q$
Molten-Native Transition IV

- Applicability to heterogeneous sequences
- Average numerically over many self-complementary random sequences
- Critical temperature found from specific heat
- Fraction of native contacts vanishes at phase transition
- Scaling plot confirms power laws predicted in the framework of the Gō-like model
DNA Hybridization I

- Same effects play a role in DNA hybridization

- Hybridization is widely used experimental method in molecular biology
  - Homology detection without sequencing
  - PCR
  - DNA chips
  - Sequencing by hybridization
  - Gene expression analysis
  - DNA computer

- Two single stranded DNA can form base pairs
  - with each other
  - with themselves

- Connect ends of the two DNA strands in Gedanken task
  → RNA structure formation

- Different applications need different experimental conditions
- Example: detection of weak homologies
- Single stranded DNA in solution form hybrid, if complementary enough
- Not complementary enough
  —→ remain single-stranded
- Existence of hybrids detected by enzyme
- Weak homology: need to reduce stringency (e.g. lower temperature)
- Problem: self-binding instead of hybrid formation
- Experimentalist has to know which phase is present
Branched Polymers

- RNA in molten phase equivalent to branched polymer

Possible forms of branched polymers:

- All results without self-avoidance agree with $Z \sim N^{-3/2} e^{\mu_0 N}$ and $R_g \sim N^{1/4}$

  $\rightarrow$ multiple branching irrelevant

  $\rightarrow$ pseudo-knots irrelevant
Denaturation I

- Description of denaturation
- Have to include spatial entropy $W(\ell) \sim \ell^{-d/2}$ of loops of $\ell$ unbound bases

- Hartree equation changes from

$$\hat{Z}(\mu)^{-1} = G_0^{-1}(\mu) - e^{-\varepsilon_0/T} \hat{Z}(\mu)$$

with $G_0^{-1} = e^\mu - 1$ to

$$\hat{Z}(\mu, k)^{-1} = G_0^{-1}(\mu, k) - e^{-\varepsilon_0/T} \int \hat{Z}(\mu, k) dk$$

with $G_0^{-1} = \mu + k^2$

- Studied by de Gennes (1968) for RNA in three dimensional space → no phase transition
• Repeat de Gennes calculation for arbitrary $d$

• Changing binding strength $\varepsilon_0$ or temperature $T$ leads to
  – no phase transition for $d < 4$
  – second order phase transition for $4 < d < 6$
  – first order phase transition for $6 < d$

• For any dimension we should get transition to self-avoiding walk for repulsive interactions

• What is missing to get phase transition in $d = 3$?
  – Stacking $\rightarrow$ arbitrarily sharp pseudo-transition
  – Self-avoidance of a single loop: $d/2 \rightarrow \nu d$
    (better but not yet enough)
  – Self-avoidance between different loops
  – Different description necessary in denatured phase, since contacts of secondary structure do not make sense for repulsive interactions
Directed Polymer Analogy

- Direct relation between RNA structure formation
  
  ![RNA Structure Diagram]

  and unbinding of a directed polymer
  
  ![Directed Polymer Diagram]


- Different sources of entropy:
  
  RNA  Binding/branching entropy  \( W \sim \ell^{-3/2} \)
  
  DP    Spatial entropy           \( W \sim \ell^{-d/2} \)

- Same critical behavior as unbinding transition in \( d = 3 \)

- Note: RNA interactions long-ranged, DP interactions local
Properties of the glass phase

- Properties of the glass phase (important for folding behavior)?
- Probe free energy $\Delta F$ of low energy, large scale excitations (droplets)
  \[ \Delta F(N) \sim N^\gamma \]
  \[ \gamma > 0 \quad \text{glass} \]
  \[ \gamma < 0 \quad \text{no glass} \]
- Pinching provides such excitations
- Just seen: $\Delta F(N) \geq [\epsilon_P + 2f_0(T)] \log N / \log 2$
- Numerically at low temperatures:
  \[ \Delta F(N) \sim a(T) \log N \]
- Glass very weak
- For practical purposes no difference between molten and glass phase
Typical Bias

- How much bias is needed?

- Numerical result for toy model: a ground state of a random sequence contains 95% Watson-Crick pairs

- A biologically useful structure must beat this threshold

- Numerical results for real RNA hairpins with different mutation rates $p$

![Graph showing the relationship between $p$ and $Q$ when $T=40^\circ C$, $N=200$.]

- Has to be compared with natural RNA sequences

- Systematic experiments necessary

- Evolution has to find very small number of good RNA sequences out of a vast amount of molten/glassy sequences

U. Gerland, RB, and T. Hwa
Finite size effects III

- Numerical verification of crossover length
- Use Vienna package to calculate thermodynamic quantities of repeated sequences using detailed free energy model
- Choose size $\langle h \rangle$ of the secondary structure as observable
- Fit sequence length dependence to

$$\langle h \rangle \sim N^{\frac{1}{2}} \left[ 1 + \left( \frac{N_0}{N} \right)^{\frac{1}{2}} + O \left( \frac{N_0}{N} \right) \right]$$

⇒ extract $N_0 \approx 7$ for $(GAC)_n$ in good agreement with prediction

- $(AU)_n$ and $(GC)_n$ numerically not feasible