Statistical mechanics of secondary structures formed by random RNA sequences

Ralf Bundschuh
The Ohio State University

Collaborator: Terence Hwa, University of California at San Diego

Outline:

- Introduction to RNA
- Uniform sequences: the molten phase
- Disorder: glass phase and glass transition
- Bias: the native phase
- Conclusions

supported by DAAD, Beckman foundation, and NSF
- RNA is heteropolymer of four different bases G, C, A, and U

- Primary structure: Sequence, e.g.,
  GCGGAUUUGAGUCAGUGGAGAGGACUCAGUUGGUCCUGUGUUCGAUCCACAGAAUUCGACCA

- 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 $S$ → partition function

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

- Partition function generated exactly by Hartree equation

→ 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:
Molten Phase I

- De Gennes (1968) proposed: start with uniform sequences \textsf{AUAAUAAUAU} \ldots or \textsf{GCGCAGCGCGC} \ldots

- Uniform attraction between any two elements of the polymer $\longrightarrow$ 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 $\Sum_{i,j}^{\mu,i,j} \longrightarrow$ 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 $\longrightarrow Z(N) \sim N^{-\theta} e^{\mu_0 N}$ with $\theta = \frac{3}{2}$

de Gennes, Biopolymers 6, 1968; Waterman, Adv. Math. Suppl. Studies 1, 1978
Molten Phase II

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

![Diagram of mountain representation](attachment:image.png)

- **one to one** correspondence: RNA secondary structures ↔ mountains
- all bases attract **equally strong**
  - counting structures ↔ counting mountains
  - free random walk in presence of a hard wall

\[
Z(N) \sim N^{-3/2} e^{\mu_0 N}
\]
\[
\langle h \rangle \sim N^{1/2}
\]
\[
R_g \sim N^{1/4}
\]
→ branched polymer
Glass Transition I

- 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 $Z^2$
  → two replicas which gain energy $\Delta \varepsilon$ for every common bond (●●●)
- 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 configuration (glass)
  ⇒ molten phase perturbatively stable
Glass Transition II

- Is glass phase stable?
- Study pinching excitations

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

Tang and Chaté, PRL 86 (2001)

- Assume molten phase is stable for all temperatures

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

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

\[ \Rightarrow \text{All pairs in this piece contribute pairing energy } \varepsilon_P \text{ in unpinched configuration} \]

- Estimate for pinching free energy:

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

- Combine two results:

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

- \( \varepsilon_P + 2f_0(T \to 0) > 0 \) (for four or more base alphabet \( \to \) next talk)

\[ \Rightarrow \text{contradiction for } T < T_* \]

\[ \Rightarrow \text{RNA cannot be in molten phase for } T < T_* \]

\[ \Rightarrow \text{different (glass) phase at low temperatures} \]
Glass Transition IV

- Properties of the glass phase (important for folding behavior)?
- Probe free energy $\Delta F$ of low energy, large scale excitations (droplets)
  
  $\Delta F(N) \sim N^\theta$
  
  $\theta > 0$  glass
  
  $\theta < 0$  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
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

- Model similar to Gō model of protein folding (Gō, J. Stat. Phys. 30, 1983)
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
RNA shows a large variety of interesting behavior.

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

RNA secondary structures offer an alternative approach to studying a variety of issues of general heteropolymer behavior.
Future work:
- Understand glass phase properties analytically
- glass-native transition
- more realistic RNA models, self-avoidance
- interactions between several molecules
- kinetics
- pseudo-knots
- tertiary structure
- biological applications (RNA finding, Huntington’s disease)
Biological function of RNA

- Biological functions:
  - **Structure** \(\rightarrow\) proteins
    - Ribosomal RNA
    - Transfer RNA
  - **Information** \(\rightarrow\) 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 AUAUAUAU... and GCGCGCGGC... need very long sequences (≥ 8000 bases, Tsunglin Liu & RB → B9.013)
  - Hairpin loops must contain at least 3 bases
  - Loss in binding energy large in hairpin loops

Hofacker et al., Monatshefte f. Chemie 125, 1994
• Naturally occurring in human DNA: \((\text{CAG})_n\) 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 (Tsunglin Liu & RB $\rightarrow$ B9.013)

$\Rightarrow$ Molten phase relevant

• Kinetics possibly important

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

- Partition function: order arbitrary structure by number of native contacts
  \[ Z(N, U_0) = \mathcal{W} + \mathcal{W} + \cdots + \mathcal{W} + \mathcal{W} + \cdots \]

- \( \mathcal{W} = W(\ell) \) = 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 (\( W \)) and full (\( Z \)) partition functions

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

  (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 \( Z(N, U_0 = 0) = Z_0(2N) \) 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 \( Q \)

In the molten phase

At the phase transition

In the native phase
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
  \[ \rightarrow \] 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
• RNA in molten phase equivalent to branched polymer

• Possible forms of branched polymers:

Zimm and Stockmeyer, J. Chem. Phys. 17, 1949
Lubensky et al., J. Physique 41, 1981
Lubensky and Isaacson, Phys. Rev. A 20, 1979
Parisi and Sourlas, Phys. Rev. Lett. 49, 1981

• All results without self-avoidance agree with $Z \sim N^{-3/2} e^{\mu_0 N}$ and $R_g \sim N^{1/4}$
  → multiple branching irrelevant
  → pseudo-knots irrelevant
Structure Size Scaling

- Characterize all phases 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:

  - native: $\langle h \rangle \sim N^{1}$
  - molten: $\langle h \rangle \sim N^{0.7}$
  - denatured: $\langle h \rangle \sim N^{1/2}$
**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
  \[ \rightarrow \text{no phase transition} \]
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 $\longrightarrow$ 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

D. Moroz and T. Hwa
• Ensemble average $\overline{Z^2}$ → 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
• 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 $\alpha = 1$
    $\longrightarrow$ marginally first order transition

• fraction of common contacts agrees well with Monte Carlo simulations

• “Large” critical disorder $\Delta \varepsilon_c$

• molten phase stable towards disorder
• Direct relation between RNA structure formation

\[ \text{RNA} + \text{RNA} + \ldots + \text{RNA} \]

and unbinding of a directed polymer

\[ \text{DP} + \text{DP} + \ldots + \text{DP} \]

\[ \text{Lipowsky, Europhys. Lett. 15, 1991} \]

• 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
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 $Q$ vs $p$ with $T=40^\circ C, N=200$](image)

U. Gerland, RB, and T. Hwa

- 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
Statistical mechanics of secondary structures formed by random RNA sequences

Ralf Bundschuh
The Ohio State University

Collaborator: Terence Hwa, University of California at San Diego

Outline:

- Introduction to RNA
- RNA phase diagram
- Possible applications

supported by DAAD (RB), Beckman foundation (TH), and NSF (TH)
Introduction I

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

- Primary structure: Sequence, e.g.,

  GCGGAUUUAGCUAGGUUGGAGAGCCACUGUGGAAUUCGAGGUGUCUGUUCGAUCCACAGAAUUCGACCA

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

- Spatial arrangement
  → tertiary structure
  (looks locally like DNA double helix)
RNA phase diagram I

- Concentrate on secondary structure

- Questions:
  What are generic properties of structures formed by random sequences?
  What are evolution or human RNA designers up against?

- Depends on external parameters

- C.f., water, ice, and vapor:

![Phase diagram](image.png)

- Three phases with vastly different properties
RNA phase diagram II

- Three important ingredients:
  - thermal fluctuations
  - sequence disorder
  - bias for native structure (biology)

- Possible two parameter phase diagram

- Phases have very different properties
  - native: molecule takes biologically meaningful structure
  - molten: many structures coexist
  - glass: molecule gets stuck in a random configuration
  - denatured: no structure at all
Questions:

- Do all of these phases really exist?
- How do we recognize which phase is present and when we change from one phase to another?

Questions not resolved in spite of

- 50 years of knowledge of DNA structure
- 10 years of computer simulations

Properties of phases for the first time determined mathematically

- Proof of existence of molten-glass phase transition
- Quantitative characterization of molten-native phase transition
Basic research

- **Physics**: basis for understanding of glassy systems in general (spin glasses, structural glasses, protein folding)
- **Biology**: basis for understanding evolution of RNA sequences

Possible practical applications

- **Identification** of RNA sequences in genomes

- **Quantitative modeling** of Huntington’s disease

- **Optimization** of experimental parameters in DNA hybridization