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
Introduction

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

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

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

- Spatial arrangement
  
  → tertiary structure
  (looks locally like DNA double helix)
- Secondary structure: Set of base pairs formed
- Pseudo-knots neglected
- Diagramatic representation

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

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

- Partition function generated exactly by Hartree equation

$$_{i\rightarrow j} = _{i\rightarrow j-1} \ + \sum_{k} _{i\rightarrow k\rightarrow j-1}$$

→ Electron in disordered medium, meanders
- Handle for analytical treatment
- $O(N^3)$ algorithm for exact partition function (McCaskill, Biopolymers 29, 1990.)
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
  \text{AUAAUAAUAAU...} \text{ or } \text{GCGCGCGCGC...}

- Uniform attraction between any two elements of the polymer \(\rightarrow\) 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 \(\sum_{i}^{j} \varepsilon_{i-j} \) 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 \(\rightarrow\) \(Z(N) \sim N^{-\theta} e^{\mu_0 N}\) with \(\theta = \frac{3}{2}\)

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

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

- **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

\[
\begin{align*}
Z(N) &\sim N^{-3/2} e^{\mu_0 N} \\
\langle h \rangle &\sim N^{1/2} \\
R_g &\sim N^{1/4}
\end{align*}
\]

\(\rightarrow\) branched polymer
Glass Transition I

- Is molten phase stable? \(\rightarrow\) 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 \(Z^2\)
  \(\rightarrow\) two replicas which gain energy \(\Delta\varepsilon\) for every common bond (\(\cdots\))
  \(7\Delta\varepsilon\)
- exactly solvable \(\rightarrow\) 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)
  \(\Rightarrow\) molten phase perturbatively stable
Glass Transition II

- 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:

\[ \Delta F(N) = -2T \log(N/2)^{-\frac{3}{2}} + T \log N^{-\frac{3}{2}} \approx \frac{3}{2} T \log N \]
• On the other hand: find piece of length
  \( \log N / \log 2 \) in first half exactly complementary
  to piece in second half

  \( \Rightarrow \) All pairs in this piece contribute pairing
  energy \( \varepsilon_P \) 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)

  \( \Rightarrow \) contradiction for \( T < T_* \)

  \( \Rightarrow \) RNA cannot be in molten phase for \( T < T_* \)

  \( \Rightarrow \) 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)

\[
\Delta F(N) \sim N^\theta \\
\theta > 0 \quad \text{glass} \\
\theta < 0 \quad \text{no glass}
\]

- 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
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
• 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 functions 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

Hofacker et al., Monatshefte f. Chemie 125, 1994

- Uniform sequences **AUUAUAUAUAU**... 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
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 (Tsunglin Liu & RB → B9.013)

⇒ Molten phase relevant

- Kinetics possibly important

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

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

- \[ \cdots = W(\ell) \] = 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
  \[ \hat{Z}(\mu; U_0) = \hat{W}(\mu) + \hat{W}(\mu) e^{-(\varepsilon_0 + U_0)/T} \hat{W}(\mu) \]
  \[ + \hat{W}(\mu) e^{-(\varepsilon_0 + U_0)/T} \hat{W}(\mu) e^{-(\varepsilon_0 + U_0)/T} \hat{W}(\mu) + \ldots \]
  (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 \( Z(N, U_0 = 0) = Z_0(2N) \) gives \( \hat{W} \)
Free energy as a function of the number of total contacts $K$ and the number of native contacts $Q$.
• 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
  \[ \rightarrow \text{remain single-stranded} \]
- 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:

  ![Diagram of branched polymers]

Lubensky et al., J. Physique 41, 1981
Lubensky and Isaacson, Phys. Rev. A 20, 1979

- 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
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:
**Denaturation**

- 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
Two Replicas I

- 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 $\rightarrow \ell^{-3/2}$ ⇒ 2 replicas $\rightarrow (\ell^{-3/2})^2 = \ell^{-6/2}$

- 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 $\alpha = 1$
    $\rightarrow$ 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
Directed Polymer Analogy

- Direct relation between RNA structure formation

\[
\begin{array}{c}
\text{RNA} \\
\text{structure formation}
\end{array}
\]

and unbinding of a directed polymer

\[
\begin{array}{c}
\text{directed polymer unbinding}
\end{array}
\]


- 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 $p$

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

$$T=40^\circ C, N=200$$

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

  ![RNA bases](image)

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

  GCGGAUUACUCACUUGGGAGACGACCAGUUGAAACUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGACCA

- **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:

- 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
Applications

- 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 **Huntigton’s disease**
  - Optimization of experimental parameters in **DNA hybridization**