Computational and physical models of RNA structure

Ralf Bundschuh

Ohio State University

July 19, 2007
Outline of all lectures

- Part I: Statistical physics of RNA
- Part II: Quantitative modeling of force-extension experiments
- Part III: RNA folding kinetics
- Part IV: microRNA target prediction
Part I

Statistical physics of RNA
Outline of part I

1. Boltzmann partition function
2. Molten RNA
3. Molten-native transition
4. Summary
Outline of part I

1. Boltzmann partition function
   - Secondary structure
   - Partition function
   - Recursion equation

2. Molten RNA

3. Molten-native transition

4. Summary
Definition of RNA secondary structure

**Definition**

An RNA secondary structure on an RNA sequence of length $N$ is a set $S$ of base pairs $(i, j)$ with $1 \leq i < j \leq N$ that fulfill the following conditions:

- Each base is involved in at most one base pair
- No pseudoknots, i.e., if $(i, j)$ and $(k, l)$ are base pairs with $i < k$, either $i < k < l < j$ or $i < j < k < l$
Diagrammatic representation

An RNA secondary structure can be represented by an arch diagram

- Every base is endpoint to at most one arch
- Two arches never cross
- One to one correspondence between arch diagrams that fulfill the above conditions and secondary structures
Energy models

- Each structure $S$ has a certain (free) energy $E[S]$
- Different models possible
  - All base pairs are equal $E[S] = \varepsilon_0 |S|$ ($\varepsilon_0 < 0$)
  - Energy associated with base pairs $E[S] = \sum_{(i,j) \in S} e(i,j)$
  - Turner parameters (nearest neighbor model):
    - energy associated with loops
      - Base pair stacks
      - Hairpin loops
      - Interior loops
      - Bulges
      - Multi-loops
Definition

The partition function of an RNA molecule with energy function $E[S]$ is given by

$$Z \equiv \sum_S e^{-\frac{E[S]}{RT}}$$

- $R = 1.987 \, \text{cal mol}^{-1} \text{K}^{-1}$
- Temperature $T$ in Kelvin $\Rightarrow RT \approx 0.6 \, \text{kcal mol}^{-1}$
- Ensemble free energy $F = -RT \ln Z$
- **Thermodynamics** completely specified if partition function is known
Calculating partition functions

Definition

Let $Z_{i,j}$ be the partition function for the RNA molecule starting at base $i$ and ending at base $j$. Denote this quantity by $Z_{i,j}$.

Observation

For the base pairing energy model the $Z_{i,j}$ obey the recursion equation

$$Z_{i,j} = Z_{i,j-1} + \sum_{k=i}^{j-1} Z_{i,k-1} e^{-\frac{e(k,j)}{RT}} Z_{k+1,j-1}$$

- Calculate from shortest to longest substrands
- $O(N^3)$ algorithm for arbitrary sequence
Outline of part I

1. Boltzmann partition function

2. Molten RNA
   - Energy model
   - z transform
   - Properties

3. Molten-native transition

4. Summary
Energetics in molten phase

Definition
In the molten phase of RNA every base can pair with every other base equally well, i.e., \( e(i, j) = \varepsilon_0 \).

Properties
- Applies to repetitive sequences: AUUAUUAUUAUUAUUAU, GCGCGCGCGCGCGC, GACGACGACGACGACGACGACGAC
- Applies to arbitrary sequences at temperatures close to denaturation on a coarse-grained level
- Minimum energy is always \( \frac{N}{2} \varepsilon_0 \)
- Structural entropy plays a major role
Molten phase partition function

Reminder

\[ Z_{i,j} = Z_{i,j-1} + \sum_{k=i}^{j-1} Z_{i,k-1} e^{-\frac{e(k,j)}{RT}} Z_{k+1,j-1} \]

Simplification

In the molten phase the partition function depends only on the length \( j - i \) of the substrand, not on \( i \) and \( j \) individually: \( Z_{i,j} \equiv G(j - i + 2) \)

Consequence

\[ G(N + 1) = G(N) + q \sum_{k=1}^{N-1} G(k) G(N - k) \quad \text{with} \quad q \equiv e^{-\frac{e_0}{RT}} \]
z transform

Definition

For any series $Q(N)$ the z transform $\hat{Q}$ is defined as

$$\hat{Q}(z) \equiv \sum_{N=1}^{\infty} Q(N)z^{-N}$$

Properties

- Function of the complex variable $z$
- Analytic outside a radius of convergence
- Discrete version of Fourier transform
- Back transform: $Q(N) = \frac{1}{2\pi i} \oint \hat{Q}(z)z^{N-1}dz$
- Convolution property: $\sum_{k=1}^{N-1} Q(k)W(N-k) = \hat{Q}(z) \cdot \hat{W}(z)$
$\hat{G}(z)$ can be calculated:
z transform for molten RNA

\( \hat{G}(z) \) can be calculated:

\[
G(N + 1) = G(N) + q \sum_{k=1}^{N-1} G(k) G(N - k)
\]
\( \hat{G}(z) \) can be calculated:

\[
G(N + 1) = G(N) + q \sum_{k=1}^{N-1} G(k) G(N - k)
\]

\[
G(N + 1)z^{-N} = G(N)z^{-N} + qz^{-N} \sum_{k=1}^{N-1} G(k) G(N - k)
\]
\[ \hat{G}(z) \text{ can be calculated:} \]

\[
G(N + 1) = G(N) + q \sum_{k=1}^{N-1} G(k)G(N - k)
\]

\[
G(N + 1)z^{-N} = G(N)z^{-N} + qz^{-N} \sum_{k=1}^{N-1} G(k)G(N - k)
\]

\[
\sum_{N=1}^{\infty} G(N + 1)z^{-N} = \hat{G}(z) + q\hat{G}^2(z)
\]
\( \hat{G}(z) \) can be calculated:

\[
G(N + 1) = G(N) + q \sum_{k=1}^{N-1} G(k)G(N - k)
\]

\[
G(N + 1)z^{-N} = G(N)z^{-N} + qz^{-N} \sum_{k=1}^{N-1} G(k)G(N - k)
\]

\[
\sum_{N=1}^{\infty} G(N + 1)z^{-N} = \hat{G}(z) + q\hat{G}^2(z)
\]

\[
z\hat{G}(z) - G(1) = \hat{G}(z) + q\hat{G}^2(z)
\]
**z transform for molten RNA**

\( \hat{G}(z) \) can be calculated:

\[
G(N + 1) = G(N) + q \sum_{k=1}^{N-1} G(k)G(N - k)
\]

\[
G(N + 1)z^{-N} = G(N)z^{-N} + qz^{-N} \sum_{k=1}^{N-1} G(k)G(N - k)
\]

\[
\sum_{N=1}^{\infty} G(N + 1)z^{-N} = \hat{G}(z) + q\hat{G}^2(z)
\]

\[
z\hat{G}(z) - G(1) = \hat{G}(z) + q\hat{G}^2(z)
\]

\[
z\hat{G}(z) - 1 = \hat{G}(z) + q\hat{G}^2(z)
\]
\( \hat{G}(z) \) can be calculated:

\[
G(N + 1) = G(N) + q \sum_{k=1}^{N-1} G(k) G(N - k)
\]

\[
G(N + 1)z^{-N} = G(N)z^{-N} + qz^{-N} \sum_{k=1}^{N-1} G(k) G(N - k)
\]

\[
\sum_{N=1}^{\infty} G(N + 1)z^{-N} = \hat{G}(z) + q\hat{G}^2(z)
\]

\[
z\hat{G}(z) - G(1) = \hat{G}(z) + q\hat{G}^2(z)
\]

\[
z\hat{G}(z) - 1 = \hat{G}(z) + q\hat{G}^2(z)
\]

\[
\hat{G}(z) = \frac{1}{2q} \left[ z - 1 - \sqrt{(z - 1)^2 - 4q} \right]
\]
Integral expression for $G(N)$

$$G(N) = \frac{1}{2\pi i} \oint \hat{G}(z) z^{N-1} \, dz$$
Integral expression for $G(N)$

\[
G(N) = \frac{1}{2\pi i} \oint \hat{G}(z) z^{N-1} dz
\]

\[
= \frac{1}{4\pi q i} \oint \left[ z - 1 - \sqrt{(z - 1)^2 - 4q} \right] z^{N-1} dz
\]
Back transform I

Integral expression for $G(N)$

\[
G(N) = \frac{1}{2\pi i} \oint \hat{G}(z)z^{N-1}dz
\]

\[
= \frac{1}{4\pi q i} \oint \left[ z - 1 - \sqrt{(z - 1)^2 - 4q} \right] z^{N-1}dz
\]

\[
= -\frac{1}{4\pi q i} \oint \sqrt{(z - 1)^2 - 4q} z^{N-1}dz
\]
Back transform I

Integral expression for $G(N)$

$$G(N) = \frac{1}{2\pi i} \oint \hat{G}(z) z^{N-1} dz$$

$$= \frac{1}{4\pi q i} \oint \left[ z - 1 - \sqrt{(z - 1)^2 - 4q} \right] z^{N-1} dz$$

$$= -\frac{1}{4\pi q i} \oint \sqrt{(z - 1)^2 - 4q} z^{N-1} dz$$

Singularity structure

- Branch cut from $1 - 2\sqrt{q}$ to $z_0 \equiv 1 + 2\sqrt{q}$
Back transform I

Integral expression for $G(N)$

$$G(N) = \frac{1}{2\pi i} \oint \hat{G}(z) z^{N-1} dz$$

$$= \frac{1}{4\pi q i} \oint \left[ z - 1 - \sqrt{(z - 1)^2 - 4q} \right] z^{N-1} dz$$

$$= -\frac{1}{4\pi q i} \oint \sqrt{(z - 1)^2 - 4q} z^{N-1} dz$$

Singularity structure

- Branch cut from $1 - 2\sqrt{q}$ to $z_0 \equiv 1 + 2\sqrt{q}$
- Contour has to go around branch cut
Back transform I

Integral expression for $G(N)$

\[
G(N) = \frac{1}{2\pi i} \oint \hat{G}(z) z^{N-1} dz
\]

\[
= \frac{1}{4\pi q i} \oint \left[ z - 1 - \sqrt{(z - 1)^2 - 4q} \right] z^{N-1} dz
\]

\[
= -\frac{1}{4\pi q i} \oint \sqrt{(z - 1)^2 - 4q} z^{N-1} dz
\]

Singularity structure

- Branch cut from $1 - 2\sqrt{q}$ to $z_0 \equiv 1 + 2\sqrt{q}$
- Contour has to go around branch cut
- Contour can be contracted to branch cut
Reminder

\[ G(N) = -\frac{1}{4\pi q_i} \oint \sqrt{(z - 1)^2 - 4qz^{N-1}}dz \]
Reminder

\[ G(N) = -\frac{1}{4\pi q i} \oint \sqrt{(z - 1)^2 - 4q z^{N-1}} \, dz \]

Leading behavior

For large \( N \) only the singularity with largest real part contributes

\[ G(N) \approx z_0^N = (1 + 2 \sqrt{q})^N \]
Reminder

\[ G(N) = -\frac{1}{4\pi q i} \oint \sqrt{(z - 1)^2 - 4q z^{N-1}} \, dz \]

Leading behavior

For large \( N \) only the singularity with largest real part contributes

\[ \Rightarrow G(N) \approx z_0^N = (1 + 2\sqrt{q})^N \]

Prefactor

\[ \int_{\mu_0}^{\mu_c} (\mu_c - \mu)^\alpha e^{\mu N} \, d\mu \approx \Gamma(1 + \alpha) N^{-(1 + \alpha)} e^{\mu_c N} \]

\[ \Rightarrow G(N) \sim N^{-3/2} (1 + 2\sqrt{q})^N \]
Properties of molten RNA

\[ G(N) \approx \left( \frac{1 + 2\sqrt{q}}{4\pi q^{3/2}} \right)^{1/2} N^{-3/2} (1 + 2\sqrt{q})^N \]

- \( N^{-3/2} \) characteristic behavior due to entropy
- Can be observed in pairing probability:

\[ \Pr\{1 \text{ and } k \text{ paired}\} = q \frac{G(k)G(N-k)}{G(N+1)} \sim k^{-3/2} \frac{(N-k)^{-3/2}}{N^{-3/2}} \approx k^{-3/2} \]

- Free energy is \( F = -RT \ln G(N) \approx -RTN \ln(1 + 2\sqrt{q}) \)
- For \( q \gg 1 \): \( F \approx -RTN \ln(2\sqrt{q}) = -\frac{RT}{2} N \ln(q) = N \frac{\varepsilon_0}{2} \)
- For \( q = 1 \): \( G(N) \approx \) number of secondary structures \( \approx N^{-3/2} 3^N \).
Outline of part I

1. Boltzmann partition function
2. Molten RNA
3. Molten-native transition
   - Model
   - Solution
   - Results
4. Summary
Model for molten-native transition

Observation
Structural RNAs have to fold into a specific “native” structure \( \Rightarrow \) there must be something in the sequence that prefers this structure

Model

- Use **perfect hairpin** as native structure

\[
\begin{array}{c}
1 \\
\quad \\
\ldots \\
\quad \\
N/2 \\
\end{array}
\begin{array}{c}
N \\
\quad \\
\ldots \\
\quad \\
1 \\
\end{array}
\]

- Assign binding energy \( \varepsilon_1 \) to all native base pairs
- Assign binding energy \( \varepsilon_0 \) to all other base pairs
Definition

Let $Z(N; q, \tilde{q})$ be the partition function of the Gō model for the molten-native transition with $2N - 2$ bases, $q = e^{-\varepsilon_0/RT}$, and $\tilde{q} = e^{-\varepsilon_1/RT}$.

Definition

Let $W(N; q) \equiv Z(N; q, \tilde{q} = 0)$ be the partition function of the Gō model for $2N - 2$ bases in which native contacts are disallowed.

Observation

$Z(N; q, \tilde{q}) = (\text{p.f. with 0 native contacts}) + (\text{p.f. with 1 native contacts}) + (\text{p.f. with 2 native contacts}) + \ldots$
Partition function

Definition

Let \( Z(N; q, \tilde{q}) \) be the partition function of the G\( \ddot{o} \) model for the molten-native transition with \( 2N - 2 \) bases, \( q = e^{-\varepsilon_0/RT} \), and \( \tilde{q} = e^{-\varepsilon_1/RT} \).

Definition

Let \( W(N; q) \equiv Z(N; q, \tilde{q} = 0) \) be the partition function of the G\( \ddot{o} \) model for \( 2N - 2 \) bases in which native contacts are disallowed.

Observation

\[
Z(N; q, \tilde{q}) = (\text{p.f. with 0 native contacts}) + (\text{p.f. with 1 native contacts}) + (\text{p.f. with 2 native contacts}) + \ldots
\]

\[
= \text{\includegraphics[width=0.5\textwidth]{nativecontacts}} + \text{\includegraphics[width=0.5\textwidth]{nativecontacts}} + \ldots + \text{\includegraphics[width=0.5\textwidth]{nativecontacts}}
\]
Individual terms
Individual terms

0 native contacts

(p.f. with 0 native contacts) $= W(N; q)$

$\Rightarrow$ z-transform: $\hat{W}(z)$
Individual terms

0 native contacts

\((\text{p.f. with 0 native contacts}) = \overset{\text{Molten-native transition}}{=} \mathcal{W}(N; q)\)

\(\Rightarrow \) \(z\)-transform: \(\hat{\mathcal{W}}(z)\)

1 native contact

\((\text{p.f. with 1 native contacts}) = \overset{\text{Solution}}{=} \tilde{q} \sum_{k=1}^{N-1} \mathcal{W}(k; q) \mathcal{W}(N - k; q)\)

\(\Rightarrow \) \(z\)-transform: \(\tilde{q} \hat{\mathcal{W}}^2(z)\)
Individual terms

0 native contacts

\[(p.f. \text{ with 0 native contacts}) = \hat{W}(N; q) \Rightarrow z\text{-transform: } \hat{W}(z)\]

1 native contact

\[(p.f. \text{ with 1 native contacts}) = \tilde{q} \sum_{k=1}^{N-1} W(k; q) W(N - k; q) \Rightarrow z\text{-transform: } \tilde{q} \hat{W}^2(z)\]

2 native contacts

\[(p.f. \text{ with 2 native contacts}) = \tilde{q}^2 \sum_{1 \leq k_1 < k_2 < N} W(k_1; q) W(k_2 - k_1; q) W(N - k_2; q) \Rightarrow z\text{-transform: } \tilde{q}^2 \hat{W}^3(z)\]
Putting it all together

Summing up

\[
\hat{Z}(z; \tilde{q}, q) = \hat{W}(z) + \tilde{q}\hat{W}(z) + \tilde{q}^2\hat{W}(z) + \cdots = \hat{E}(z)
\]

What is $\hat{W}(z)$?

For $\tilde{q} = q$ we have

\[
Z(N; q, \tilde{q} = q) = G(2N - 1; q)
\]

$\Rightarrow$

\[
\hat{Z}(z; q, q) = \hat{E}(z) \equiv \sum_{N=1}^{\infty} G(2N - 1; q)z^{-N}
\]

$\Rightarrow$

\[
\hat{W}(z) = \hat{E}(z)
\]

$\Rightarrow$

\[
\hat{Z}(z; q, \tilde{q}) = \hat{E}(z)
\]
Putting it all together

Summing up

$\hat{Z}(z; \tilde{q}, q) = \hat{W}(z) + \tilde{q}\hat{W}^2(z) + \tilde{q}^2\hat{W}^3(z) + \ldots = \hat{E}(z)$
Putting it all together

Summing up

\[
\hat{Z}(z; \tilde{q}, q) = \hat{W}(z) + \tilde{q}\hat{W}^2(z) + \tilde{q}^2\hat{W}^3(z) + \ldots = \frac{\hat{W}(z)}{1 - \tilde{q}\hat{W}(z)}
\]
Putting it all together

Summing up

\[ \hat{Z}(z; \tilde{q}, q) = \hat{W}(z) + \tilde{q} \hat{W}^2(z) + \tilde{q}^2 \hat{W}^3(z) + \ldots = \frac{\hat{W}(z)}{1 - \tilde{q} \hat{W}(z)} \]

What is \( \hat{W}(z) \)?
Putting it all together

Summing up

\[ \hat{Z}(z; \tilde{q}, q) = \hat{W}(z) + \tilde{q} \hat{W}^2(z) + \tilde{q}^2 \hat{W}^3(z) + \ldots = \frac{\hat{W}(z)}{1 - \tilde{q} \hat{W}(z)} \]

What is \( \hat{W}(z) \)?

For \( \tilde{q} = q \), we have

\[ Z(N; q, \tilde{q} = q) = G(2N - 1; q) \]
Putting it all together

Summing up

\[
\hat{Z}(z; \tilde{q}, q) = \hat{W}(z) + \tilde{q}\hat{W}^2(z) + \tilde{q}^2\hat{W}^3(z) + \ldots = \frac{\hat{W}(z)}{1 - \tilde{q}\hat{W}(z)}
\]

What is \(\hat{W}(z)\)?

For \(\tilde{q} = q\) we have \(Z(N; q, \tilde{q} = q) = G(2N - 1; q)\)

\[\Rightarrow \hat{Z}(z; q, q) = \hat{E}(z; q) \equiv \sum_{N=1}^{\infty} G(2N - 1; q)z^{-N}\]
Putting it all together

Summing up

\[
\hat{Z}(z; \tilde{q}, q) = \hat{W}(z) + \tilde{q}\hat{W}^2(z) + \tilde{q}^2\hat{W}^3(z) + \ldots = \frac{\hat{W}(z)}{1 - \tilde{q}\hat{W}(z)}
\]

What is \( \hat{W}(z) \)?

For \( \tilde{q} = q \) we have

\[
Z(N; q, \tilde{q} = q) = G(2N - 1; q)
\]

\[
\Rightarrow \hat{Z}(z; q, q) = \hat{E}(z; q) \equiv \sum_{N=1}^{\infty} G(2N - 1; q)z^{-N}
\]

\[
\Rightarrow \hat{W}(z) = \frac{\hat{E}(z)}{1 + q\hat{E}(z)}
\]
Putting it all together

Summing up

\[ \hat{Z}(z; \tilde{q}, q) = \hat{W}(z) + \tilde{q}\hat{W}^2(z) + \tilde{q}^2\hat{W}^3(z) + \ldots = \frac{\hat{W}(z)}{1 - \tilde{q}\hat{W}(z)} \]

What is \( \hat{W}(z) \)?

For \( \tilde{q} = q \) we have \( Z(N; q, \tilde{q} = q) = G(2N - 1; q) \)

\[ \Rightarrow \hat{Z}(z; q, q) = \hat{E}(z; q) \equiv \sum_{N=1}^{\infty} G(2N - 1; q)z^{-N} \]

\[ \Rightarrow \hat{W}(z) = \frac{\hat{E}(z)}{1 + q\hat{E}(z)} \quad \Rightarrow \quad \hat{Z}(z; q, \tilde{q}) = \frac{\hat{E}(z)}{1 - (\tilde{q} - q)\hat{E}(z)} \]
Behavior of $\hat{E}(z)$

Expression for $\hat{E}(z)$

$$\hat{E}(z) = \sum_{N=1}^{\infty} G(2N - 1; q) z^{-N}$$

can be calculated similarly to $\hat{G}(z)$.

$$\hat{E}(z) = \frac{1}{2q} - \frac{1}{4qz} \left[ \sqrt{(z - 1)^2 - 4q} + \sqrt{(z + 1)^2 - 4q} \right].$$

Properties

- Vanishes as $z \to \infty$
- Square root branch cut at $z_0 = 1 + 2\sqrt{q}$
- Finite limit $\hat{E}(1 + 2\sqrt{q})$ at branch cut
- Other branch cuts have smaller real part
Singularity structure of $\hat{Z}(z)$

Candidate singularities

$$\hat{Z}(z; q, \tilde{q}) = \frac{\hat{E}(z)}{1 - (\tilde{q} - q)\hat{E}(z)}$$

- Square root branch cut at $z_0 = 1 + 2\sqrt{q}$
- Pole at $z_1(\tilde{q})$ given by $\hat{E}(z_1) = 1/(\tilde{q} - q)$

Dominant singularity

- Square root branch cut at $z_0$ if $1/(\tilde{q} - q) > \hat{E}(z_0)$
- Pole at $z_1(\tilde{q})$ if $1/(\tilde{q} - q) < \hat{E}(z_0)$

$\Rightarrow$ Phase transition
Properties of molten-native transition

Characterization of phase transition

- Critical bias $\tilde{q}_c = q + 1/\hat{E}(z_0)$
- If $\tilde{q} < \tilde{q}_c$ regular molten behavior $Z \sim N^{-3/2}(1 + 2\sqrt{\tilde{q}})^N$
  $\longrightarrow$ native base pairs do not play any role
- If $\tilde{q} > \tilde{q}_c$ native behavior with $Z \sim N^0 z_1(\tilde{q})^N$
  $\longrightarrow$ finite fraction of native base pairs but still many “bubbles”
- For $\tilde{q} \gg \tilde{q}_c$ we get $Z \sim N^0 \tilde{q}^N$
  $\longrightarrow$ only native base pairs

Conclusion

It takes a finite amount of sequence bias to enforce a native structure.
Outline of part I

1. Boltzmann partition function
2. Molten RNA
3. Molten-native transition
4. Summary
Summary of part I

- The partition function of RNA can be calculated in \textit{polynomial time}.
- Asymptotic behavior for homogeneous RNA can be calculated by analytical methods.
- The partition function of molten RNA has a \textit{characteristic} $N^{-3/2}$ behavior.
- It takes a \textit{finite amount of sequence bias} to enforce a native structure.