Computational and physical models of RNA structure

Ralf Bundschuh

Ohio State University

July 25, 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

Outline of part I

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

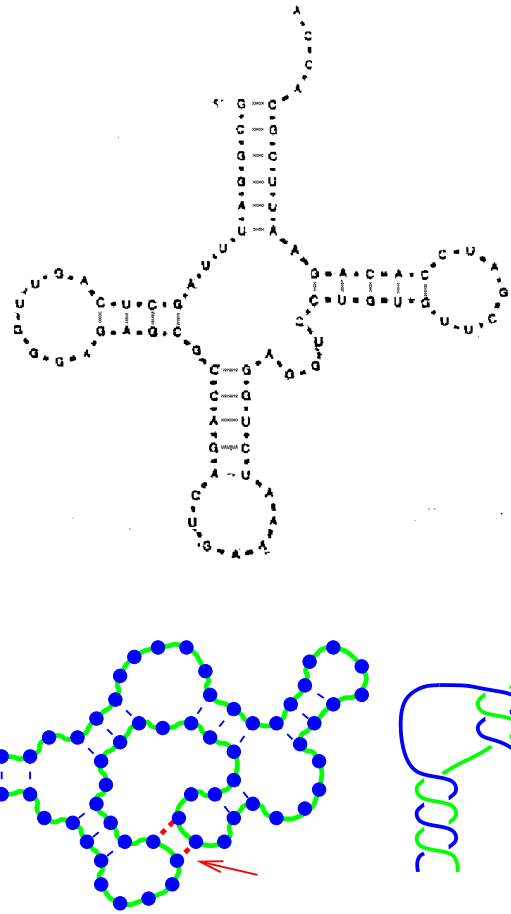
Outline of part I

Definition of RNA secondary structure

Definition

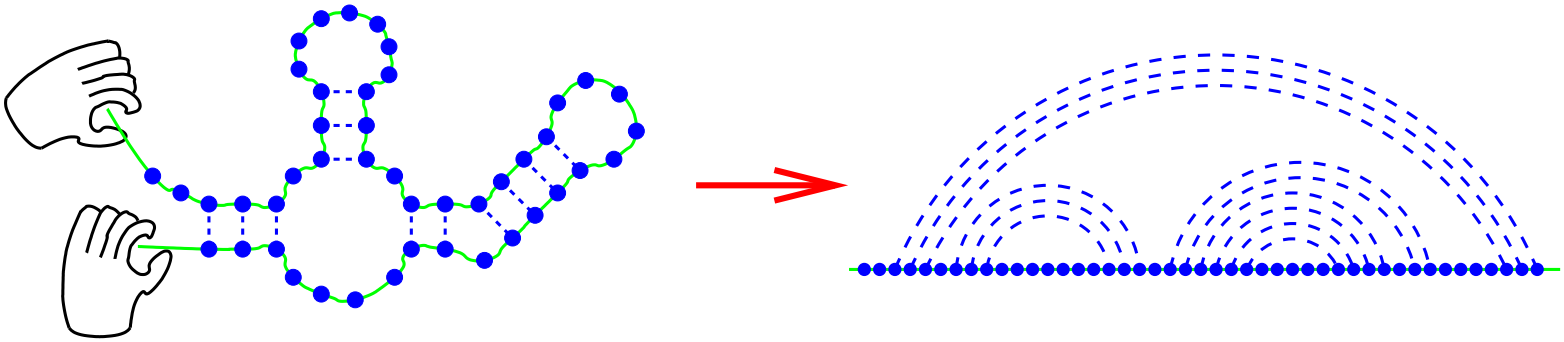
An RNA secondary structure on an RNA sequence of length N is a set S of base pairs (i, j) with $1 \leq i < j \leq N$ that fulfill the following conditions:

- 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

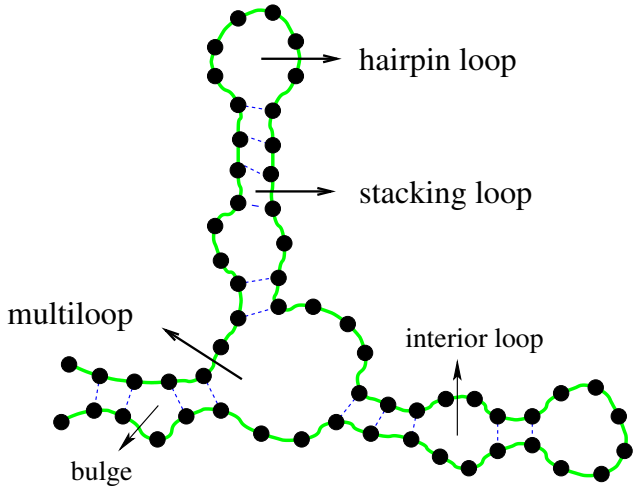
An RNA secondary structure can be represented by an arch diagram



- Every base is end point to at most one arch
- Two arches never cross
- One to one correspondance between arch diagrams that fulfil 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 $E[S] = \varepsilon_0 |S| \qquad (\varepsilon_0 < 0)$
  - 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



Partition function

Definition

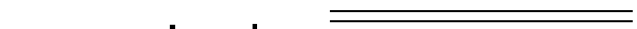
The partition function of an RNA molecule with energy function $E[S]$ is given by

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

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

Calculating partition functions

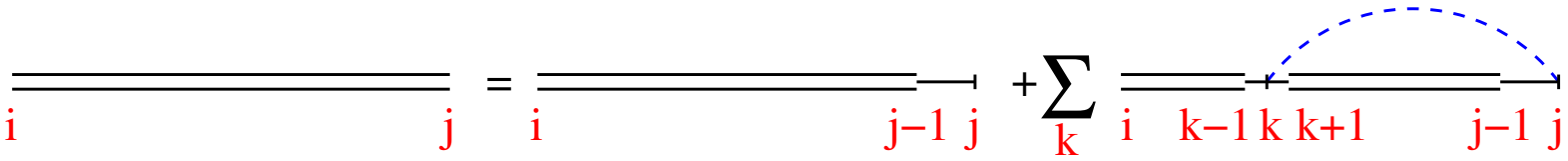
Definition

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 

.

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

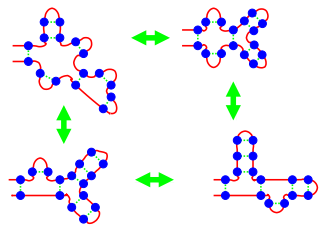
Energetics in molten phase

Definition

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

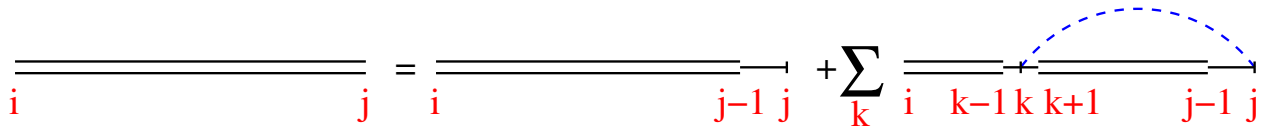
Properties

- Applies to repetitive sequences: `AUAUAUAUAUAUAUAUAU`, `GCGCGCGCGCGCGC`, `GACGACGACGACGACGACGAC`
- 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

$$G(N+1) = G(N) + q \sum_{k=1}^{N-1} G(k)G(N-k) \qquad \text{with} \quad q \equiv e^{-\frac{\varepsilon_0}{RT}}$$

z transform

Definition

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

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

Properties

- Function of the complex variable z
- Analytic outside a radius of convergence
- Discrete version of Fourier transform
- Back transform: $Q(N) = \frac{1}{2\pi i} \oint \widehat{Q}(z) z^{N-1} \mathrm{d}z$
- Convolution property: $\widehat{\sum_{k=1}^{N-1} Q(k) W(N-k)} = \widehat{Q}(z) \cdot \widehat{W}(z)$

z transform for molten RNA

$\widehat{G}(z)$ can be calculated:

$$
\begin{aligned}
G(N+1) &= G(N) + q\sum_{k=1}^{N-1} G(k)G(N-k) \\
G(N+1)z^{-N} &= G(N)z^{-N} + qz^{-N}\sum_{k=1}^{N-1} G(k)G(N-k) \\
\sum_{N=1}^{\infty} G(N+1)z^{-N} &= \widehat{G}(z) + q\widehat{G}^2(z) \\
z\widehat{G}(z) - G(1) &= \widehat{G}(z) + q\widehat{G}^2(z) \\
z\widehat{G}(z) - 1 &= \widehat{G}(z) + q\widehat{G}^2(z)
\end{aligned}
$$

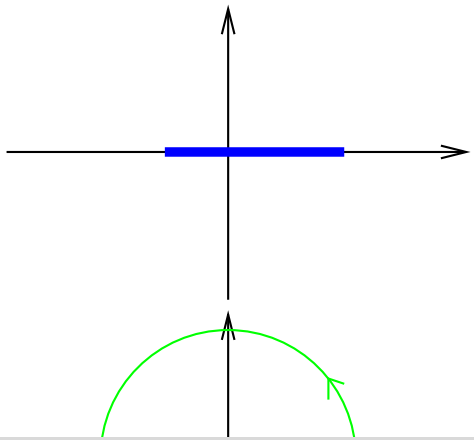
$$
\boxed{\widehat{G}(z) = \frac{1}{2q}\left[z - 1 - \sqrt{(z-1)^2 - 4q}\right]}
$$

Back transform I

Integral expression for $G(N)$

$$\begin{aligned} G(N) &= \frac{1}{2\pi i}\oint \widehat{G}(z) z^{N-1}\mathrm{d}z \\ &= \frac{1}{4\pi q i}\oint \left[z-1-\sqrt{(z-1)^2-4q}\right] z^{N-1}\mathrm{d}z \\ &= -\frac{1}{4\pi q i}\oint \sqrt{(z-1)^2-4q}\, z^{N-1}\mathrm{d}z \end{aligned}$$

Singularity structure

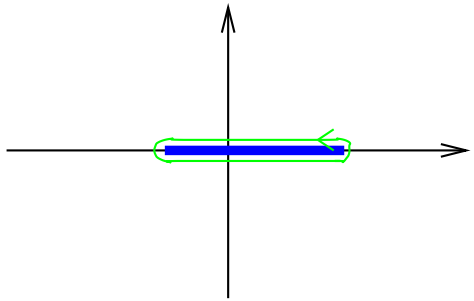


- Branch cut from $1-2\sqrt{q}$ to $z_0 \equiv 1+2\sqrt{q}$

Back transform II

Reminder

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



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 $q = 1$: $G(N) =$number of secondary structures$\sim N^{-3/2}3^N$.

Outline of part I

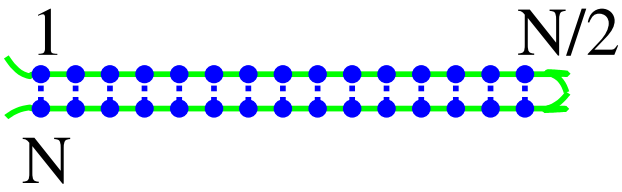
Model for molten-native transition

Observation

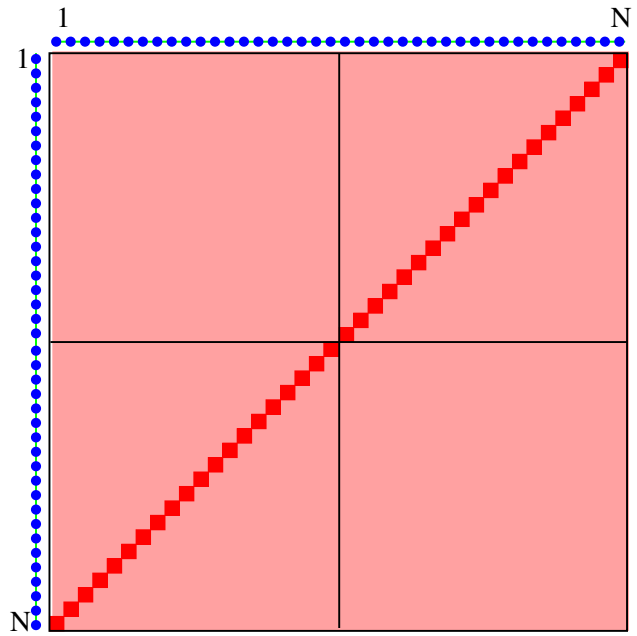
Structural RNAs have to fold into a specific "native" structure
$\Rightarrow$ there must be something in the sequence that prefers this structure

Model

- Use perfect hairpin as native structure



- Assign binding energy ε_1 to all native base pairs
- Assign binding energy ε_0 to all other base pairs



Partition function

Definition

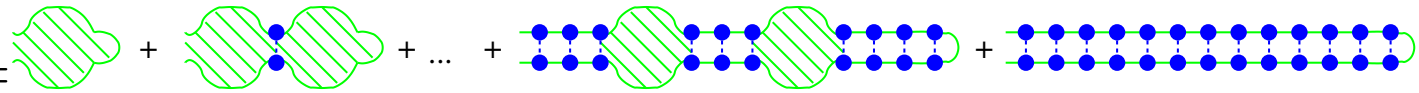
Let $Z(N; q, \widetilde{q})$ be the partition function of the Gō model for the molten-native transition with $2N-2$ bases, $q = e^{-\varepsilon_0/RT}$, and $\widetilde{q} = e^{-\varepsilon_1/RT}$

Definition

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

Observation

$Z(N; q, \widetilde{q}) =$(p.f. with 0 native contacts) + (p.f. with 1 native contacts) + (p.f. with 2 native contacts) + ...

$=$ 

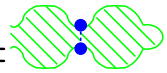
Individual terms

0 native contacts

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

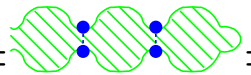
$\Rightarrow$ z-transform: $\widehat{W}(z)$

1 native contact

(p.f. with 1 native contacts) $=$  $= \widetilde{q}\sum_{k=1}^{N-1} W(k;q)W(N-k;q)$

$\Rightarrow$ z-transform: $\widetilde{q}\widehat{W}^2(z)$

2 native contacts

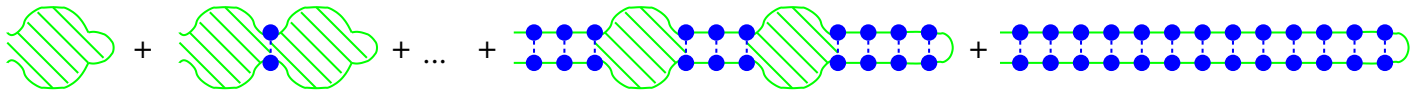
(p.f. with 2 native contacts) $=$  $=$

$\widetilde{q}^2\sum_{1\le k_1<k_2<N} W(k_1;q)W(k_2-k_1;q)W(N-k_2;q)$

$\Rightarrow$ z-transform: $\widetilde{q}^2\widehat{W}^3(z)$

Putting it all together

Summing up



$$\widehat{Z}(z;\widetilde{q},q) = \widehat{W}(z) + \widetilde{q}\widehat{W}^2(z) + \widetilde{q}^2\widehat{W}^3(z) + \ldots = \frac{\widehat{W}(z)}{1-\widetilde{q}\widehat{W}(z)}$$

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

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

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

$$\Rightarrow \widehat{W}(z) = \frac{\widehat{E}(z)}{1+q\widehat{E}(z)} \qquad \Rightarrow \boxed{\widehat{Z}(z;q,\widetilde{q}) = \frac{\widehat{E}(z)}{1-(\widetilde{q}-q)\widehat{E}(z)}}$$

Behavior of $\widehat{E}(z)$

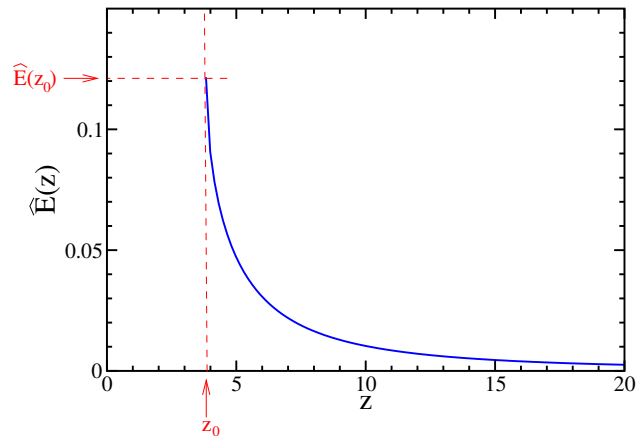
Expression for $\widehat{E}(z)$

$\widehat{E}(z) = \sum_{N=1}^{\infty} G(2N-1;q) z^{-N}$ can be calculated similarly to $\widehat{G}(z)$.

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

Properties

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

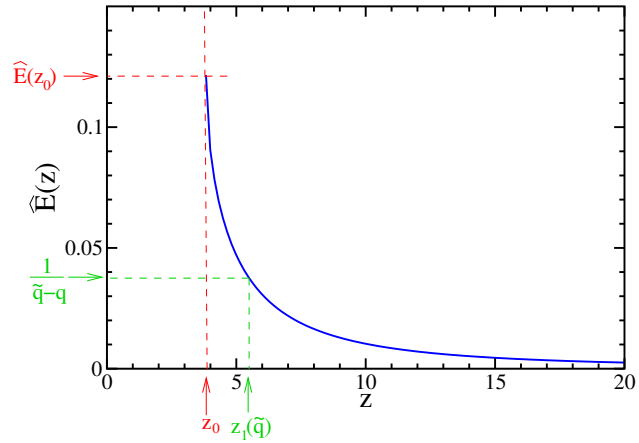


Singularity structure of $\widehat{Z}(z)$

Candidate singularities

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

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



Dominant singularity

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

$\Rightarrow$ Phase transition

Properties of molten-native transition

Characterization of phase transition

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

Conclusion

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

Outline of part I

Summary of part I

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

Outline of part II

Outline of part II

Experimental methods to determine RNA structure

Experimental methods

- X-ray crystallography
- Nuclear Magnetic Resonance
- Biochemical evidence: protection essays
- Correlated mutation analysis

Problem

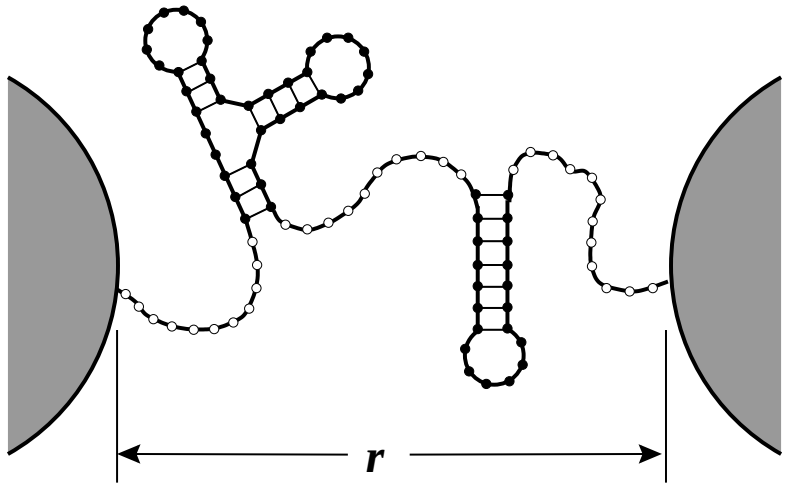
RNA is very floppy $\Rightarrow$ two RNA molecules rarely have the same three-dimensional structure

Idea

Look at one RNA molecule at a time.

Experimental setup

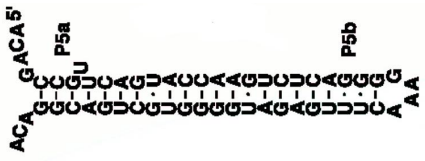
- Attach ends of RNA molecule to two beads
- Keep beads at fixed distance R with optical tweezers
- Measure force f on beads as function of distance R



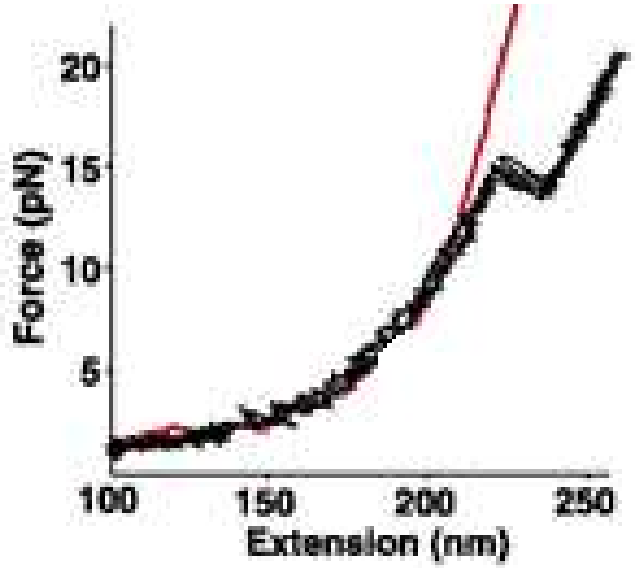
Measurement

Subject

P5ab hairpin of *Tetrahymena thermophila* group I intron



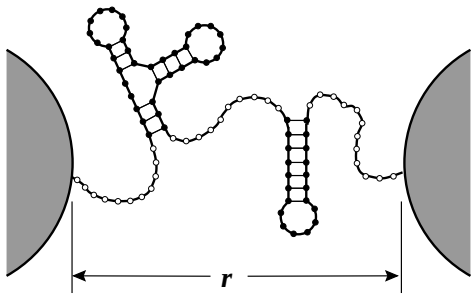
Data



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

Outline of part II

Calculating the force



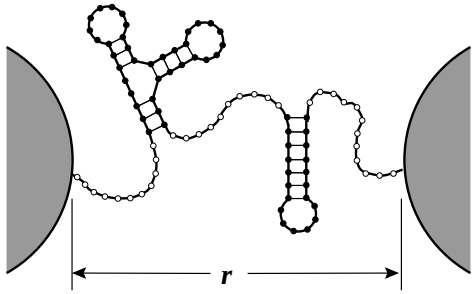
Basic physics

energy = force · distance

Force from free energy

Need free energy $F(r)$ at fixed end to end distance $r \Rightarrow$ force $f = \dfrac{\partial F(r)}{\partial r}$

Partition function I



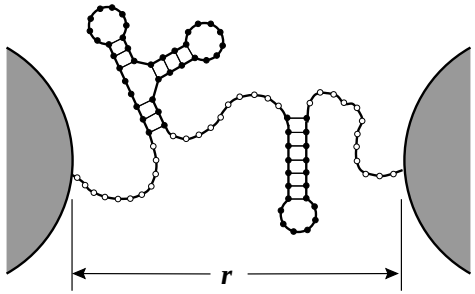
Free energy from partition function

$F(r) = -RT \ln Z(r) \Rightarrow$ need partition function $Z(r)$ at fixed distance r

Partition function

$$Z(r) = \sum_{\text{secondary structures } S} \sum_{\substack{\text{polymer configurations } \mathcal{P} \text{ with} \\ \text{distance } r \text{ and structure } S}} e^{-\frac{E[S]+E[\mathcal{P}]}{RT}}$$

Partition function II

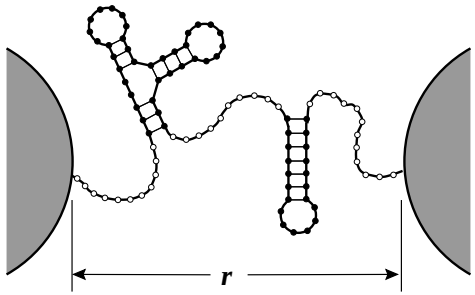


$$Z(r) = \sum_{\text{secondary structures } S} \sum_{\substack{\text{polymer configurations } \mathcal{P} \text{ with} \\ \text{distance } r \text{ and structure } S}} e^{-\frac{E[S]+E[\mathcal{P}]}{RT}}$$

$$= \sum_{\text{secondary structures } S} \sum_{\substack{\text{polymer configurations } \mathcal{P} \\ \text{of the ``exterior bases'' with} \\ \text{distance } r \text{ and structure } S}} e^{-\frac{E[S]+E[\mathcal{P}]}{RT}}$$

$$= \sum_{m=0}^{N} \sum_{\substack{\text{secondary structures } S \\ \text{with } m \text{ exterior bases}}} \sum_{\substack{\text{polymer configurations } \mathcal{P} \\ \text{of the ``exterior bases'' with} \\ \text{distance } r \text{ and structure } S}} e^{-\frac{E[S]+E[\mathcal{P}]}{RT}}$$

Partition function III



$$Z(r) = \sum_{m=0}^{N} \sum_{\substack{\text{secondary structures } S \\ \text{with } m \text{ exterior bases}}} \sum_{\substack{\text{polymer configurations } \mathcal{P} \\ \text{of the ``exterior bases'' with} \\ \text{distance } r \text{ and structure } S}} e^{-\frac{E[S]+E[\mathcal{P}]}{RT}}$$

$$= \sum_{m=0}^{N} \sum_{\substack{\text{secondary structures } S \\ \text{with } m \text{ exterior bases}}} e^{-\frac{E[S]}{RT}} \sum_{\substack{\text{polymer configurations } \mathcal{P} \\ \text{of } m \text{ ``exterior bases''} \\ \text{with distance } r}} e^{-\frac{E[\mathcal{P}]}{RT}}$$

$$= \sum_{m=0}^{N} \left(\sum_{\substack{\text{secondary structures } S \\ \text{with } m \text{ exterior bases}}} e^{-\frac{E[S]}{RT}} \right) \left(\sum_{\substack{\text{polymer configurations } \mathcal{P} \text{ of ssRNA} \\ \text{with } m \text{ bases with distance } r}} e^{-\frac{E[\mathcal{P}]}{RT}} \right)$$

Partition function IV

$$\boxed{Z(r) = \sum_{m=0}^{N} Q(m) W(m, r)}$$

with

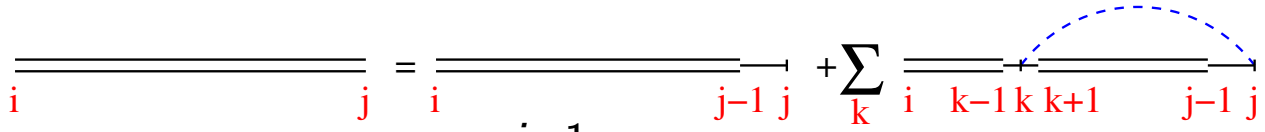
$$Q(m) \equiv \sum_{\substack{\text{secondary structures } S \\ \text{with } m \text{ exterior bases}}} e^{-\frac{E[S]}{RT}}$$

and

$$W(m, r) \equiv \sum_{\substack{\text{polymer configurations } \mathcal{P} \text{ of ssRNA} \\ \text{with } m \text{ bases with distance } r}} e^{-\frac{E[\mathcal{P}]}{RT}}$$

Recursion equation

Reminder



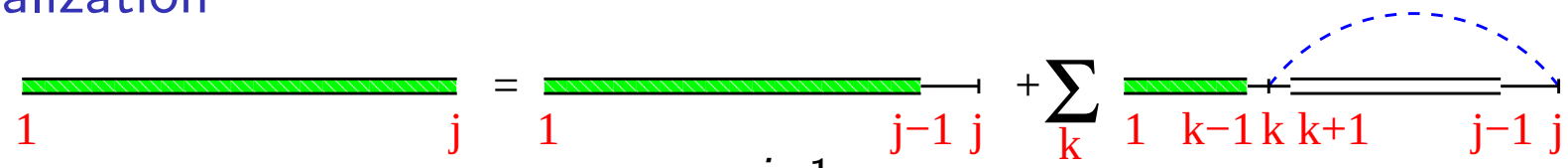
$$Z_{i,j} = Z_{i,j-1} + \sum_{k=i}^{j-1} Z_{i,k-1} e^{-\frac{e(k,j)}{RT}} Z_{k+1,j-1}$$

Definition

Let $Q_j(m)$ be the partition function for the first j bases with exactly m exterior bases (

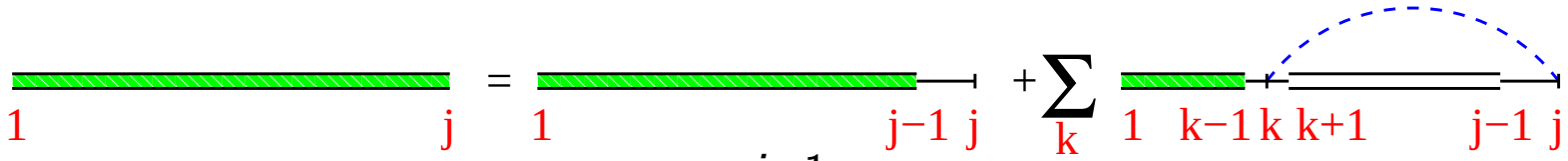
).

Generalization



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

Properties of recursion equation



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

Properties:

- $O(N^3)$ complexity
- $Q(m) = Q_N(m)$
- Can be easily generalized to Turner parameters
- Can be calculated by modifying the Vienna package (Hofacker, Fontana, Stadler, Bonhoeffer, Tacker, and Schuster, Monatshefte f. Chemie, 1994)

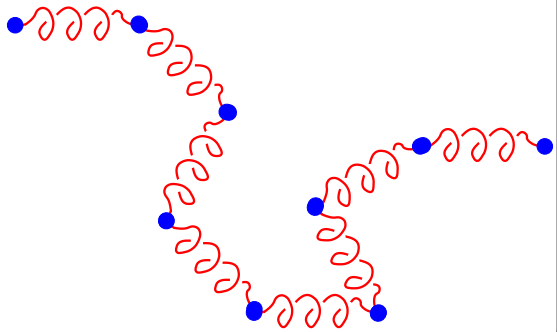
Polymer physics

Need

$$W(m,r) = \sum_{\substack{\text{polymer configurations } \mathcal{P} \text{ of ssRNA} \\ \text{with } m \text{ bases with distance } r}} e^{-\frac{E[\mathcal{P}]}{RT}}$$

Model

- Elastic freely jointed chain
- Persistence length $1.9nm$/base distance $0.7nm$



Partition function

Calculation of $W(m,r)$ is standard polymer physics.

RNApull

Putting it together

$Q(m),\ W(m,r)$

$$\longrightarrow Z(r) = \sum_{m=0}^{N} Q(m)W(m,r)$$

$$\longrightarrow F(r) = -RT \ln Z(r)$$

$$\longrightarrow f(r) = \frac{\partial F(r)}{r}$$

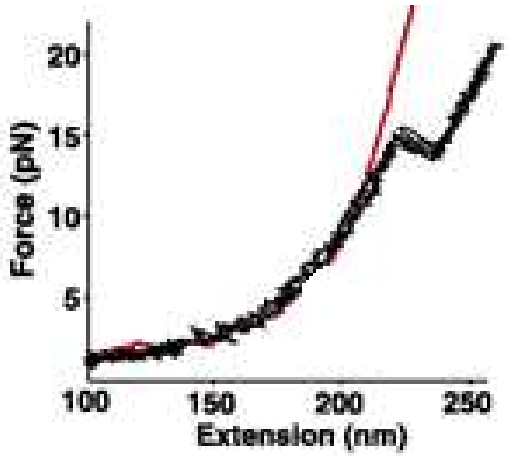
Web server

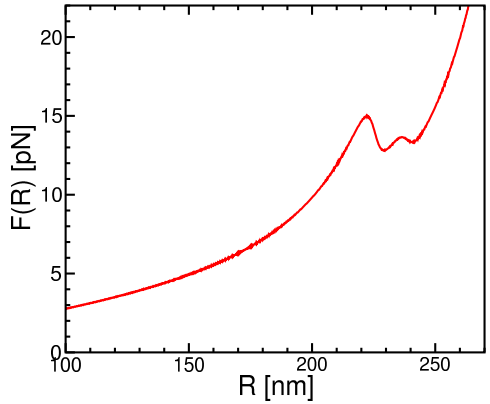
`http://bioserv.mps.ohio-state.edu/rna`

Outline of part II

Hairpin

Apply to P5ab hairpin of *Tetrahymena thermophila* group I intron



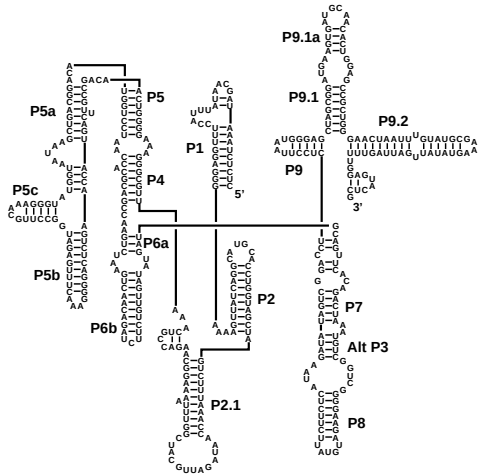


Quantitative agreement!

The full group I intron

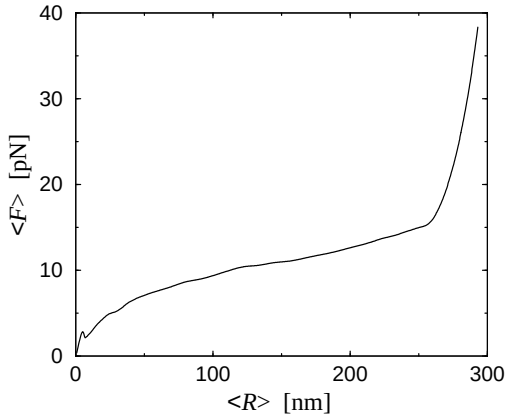
The group I intron of *Tetrahymena thermophila*

- Group I intron contains pseudo-knot!
- Quantitative modeling ignores pseudo-knot
  $\Rightarrow$ known inactive conformation
  (Pan and Woodson, J. Mol. Biol., 1998)



Computational result

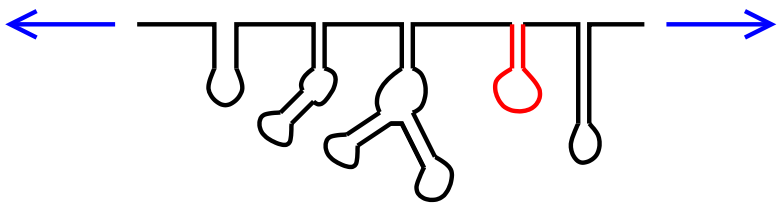
No sign of secondary structure!

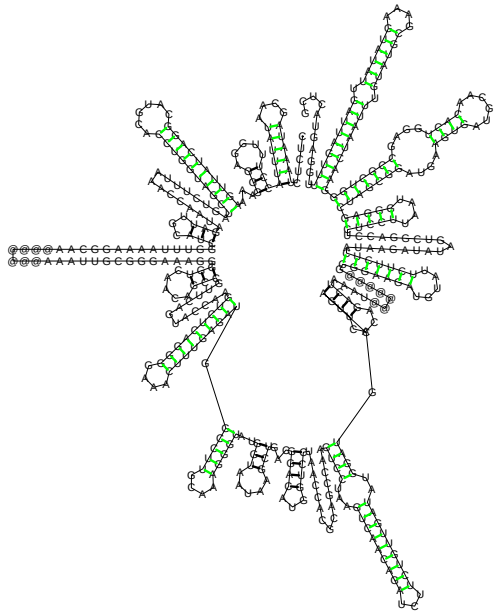


What's happening?

Intermediate structure

- Look at intermediate structure (here $r = 100$nm)
- Like "Socks on the clothes line"

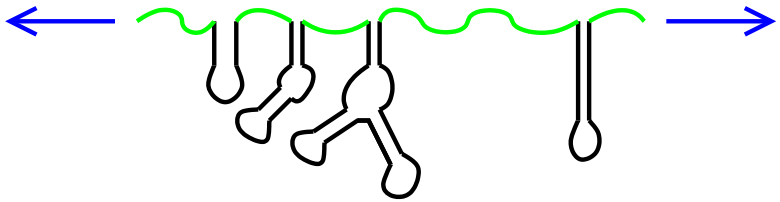




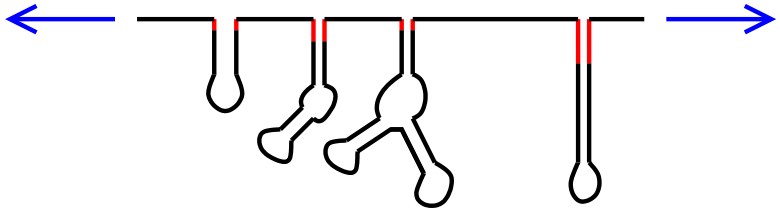
What's happening?

Compensation effect

- Extension r is increased
- One of the "socks" disappears



- The other "socks" take up the slag



$\Rightarrow$ No rapid change in force as sock disappears

$\Rightarrow$ smooth force-extension curve

Outline of part II

Nanopores

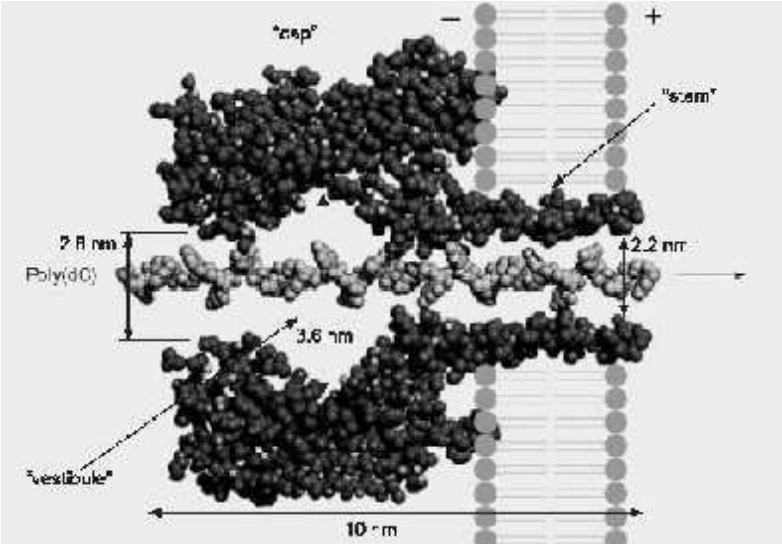
What is a nanopore?

A nanopore

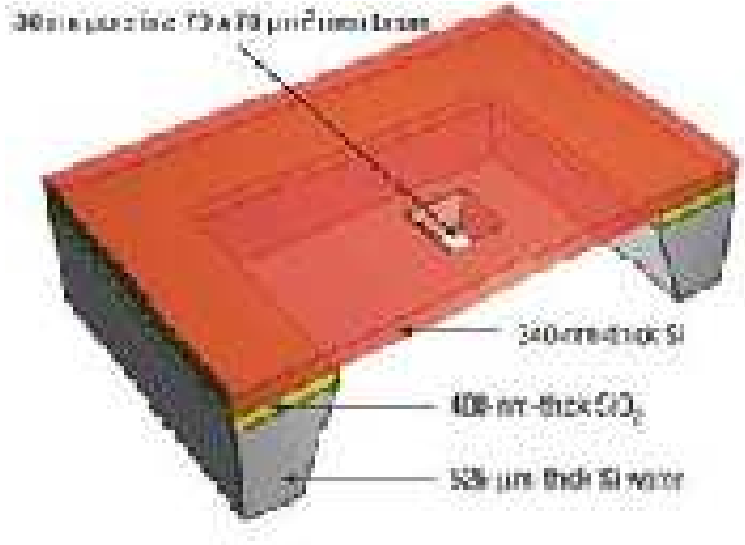
- is a little hole
- is so small that only single-stranded but no double-stranded RNA can pass through it
- can be placed between two chambers such that there is only one nanopore connecting the chambers

Nanopores

Types of nanopores



natural ion channel (α-hemolysin)
(Meller, J. Phys. Cond. Mat., 2003)

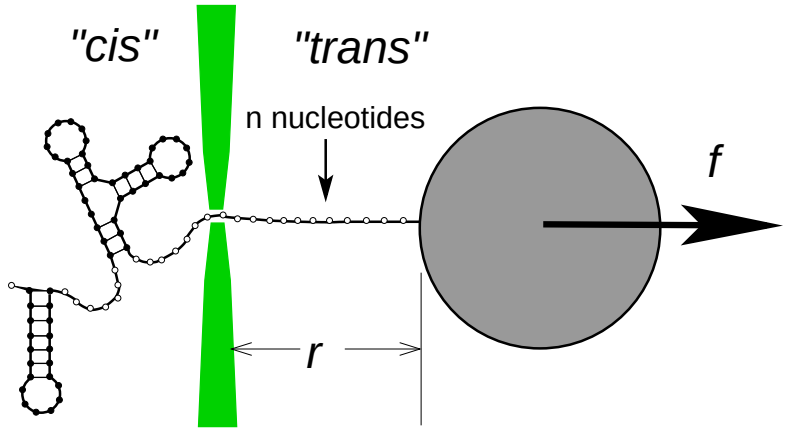


solid state pore
(Storm et al., Nature Mat., 2003)

Combining nanopores with force-extension experiments

Suggestion

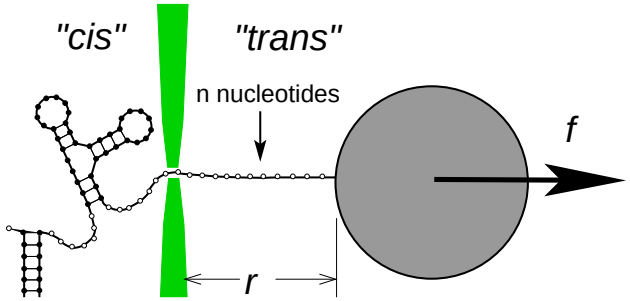
Combine a nanopore with a force-extension setup



Prediction

- The force will rise at every structural element
- The structural elements will break in their order along the sequence

Simulation approach

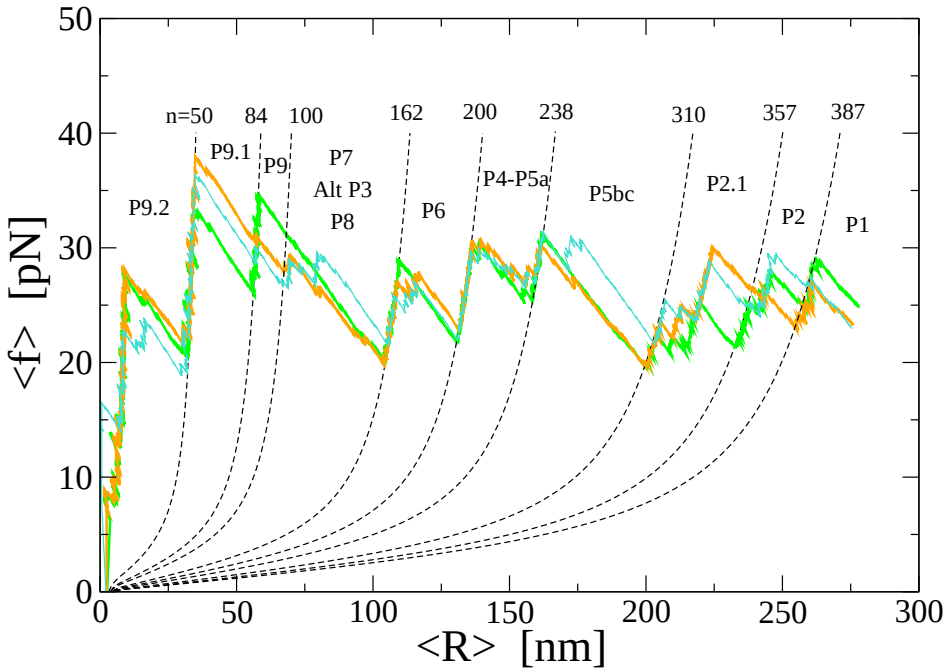


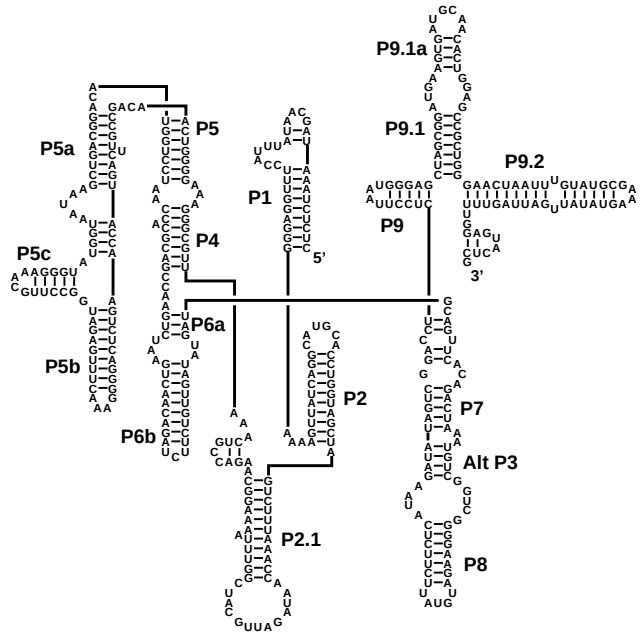
Simulation

- Fix $r(t)$ to be linear (set by experiment)
- Degree of freedom: number n of base in the pore
- At each time n can
  - increase:
    $\longrightarrow$ calculable gain in mechanical energy on the right
    $\longrightarrow$ potentially loss in binding energy calculable from $E[S]$
  - decrease:
    $\longrightarrow$ calculable loss in mechanical energy on the right
- Monte Carlo simulation (see also part III)

Simulation result

Apply to group I intron of *Tetrahymena thermophila*

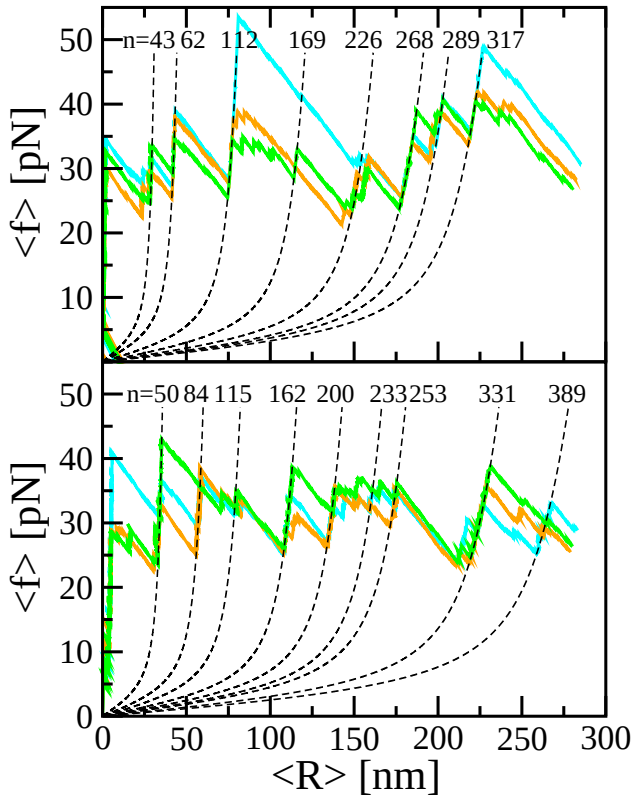




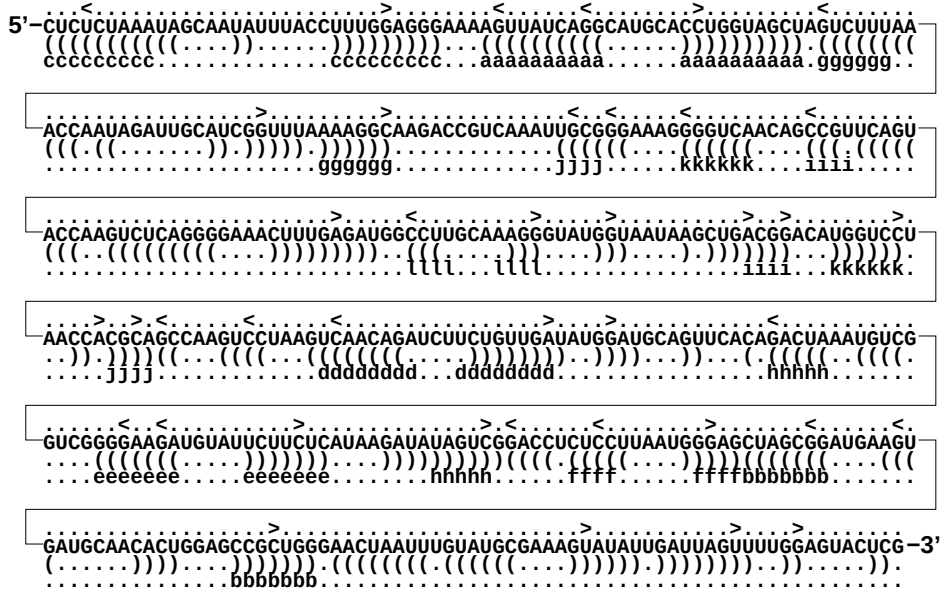
- Signature of every structural element
- Can extract sequence position of stalling sites

Structure reconstruction I

Repeat pulling in opposite direction

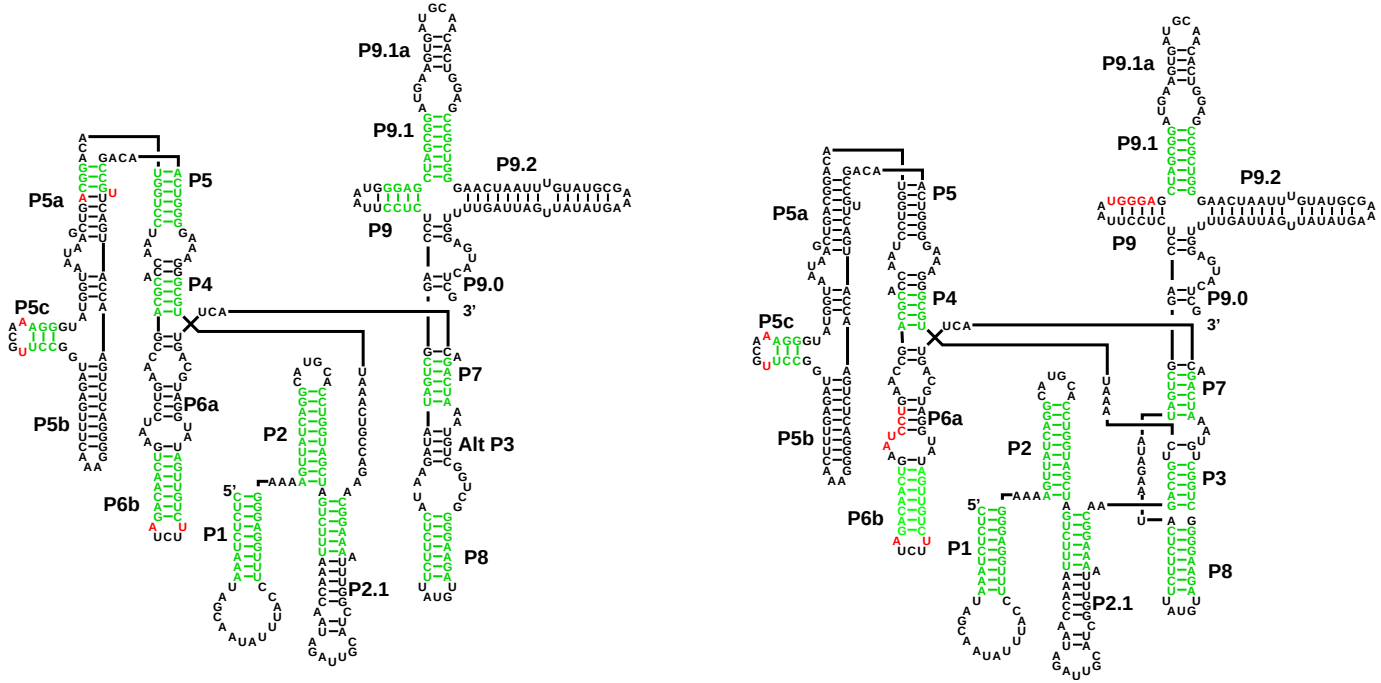


⇒



Match stalling sites by sequence comparison ⇒ reconstruct structure

Structure reconstruction II



- Overall structure reconstructed correctly
- Can distinguish different structures on the same sequence
- Even pseudoknot reconstructed
- "Just" needs to be implemented experimentally ...

Outline of part II

Summary of part II

- Quantitative description of single-molecule experiments possible
- Force-extension curves do not reveal secondary structure information due to compensation effects
- Single-molecule experiments reveal structure information with the help of a nanopore

Outline of part III

Outline of part III

Kinetics introduction

Thermodynamics

What structure(s) does an RNA molecule take on?

Kinetics

- How long does it take an RNA molecule to reach a certain structure?
- Along which pathway does an RNA molecule get to a certain structure?

Why is kinetics important?

Why is kinetics a problem?

- Take, say, a short RNA with 50 bases
- Has roughly $2.6^{50} = 6 \cdot 10^{20}$ structures
- Let's be generous and say that it can explore a structure per picosecond
  $\Rightarrow$ it takes 500,000,000s$\approx$ 18yr to explore all structures!

Processes that need to happen in time:

- rRNA have to fold
- riboswitches have to get from one configuration to another
- terminator hairpins have to form in time for termination
- RNA viruses have to be folded for packaging
- splicing might depend on secondary structure

Outline of part III

What is Molecular Dynamics?

- Model an RNA molecule atom by atom
- Add water and salt atom by atom or as an effective medium
- Choose a force field
- Integrate Newton's equations

Advantages and disadvantages of Molecular Dynamics

Advantages

- First principle based modeling
- Full three dimensional structure included

Disadvantages

- Need many atoms ($\approx$ 30 atoms per base)
- Can only simulate very short sequences (10 base pairs)
- Can only simulate very short times (at most μs)

What can be done with MD?

Structural dynamics of very small building blocks:

- Sorin *et al.*, RNA simulations: probing hairpin unfolding and the dynamics of a GNRA tetraloop, J. Mol. Biol. 2002
- Sarzynska *et al.*, Effects of base substitutions in an RNA hairpin from molecular dynamics and free energy simulations, Biophys J. 2003
- Hart *et al.*, Molecular dynamics simulations and free energy calculations of base flipping in dsRNA, RNA. 2005
- Mazier *et al.*, Molecular dynamics simulation for probing the flexibility of the 35 nucleotide SL1 sequence kissing complex from HIV-1Lai genomic RNA, J Biomol Struct Dyn. 2007
- Nystrom *et al.*,Molecular dynamics study of intrinsic stability in six RNA terminal loop motifs, J Biomol Struct Dyn. 2007

Outline of part III

What is Monte Carlo dynamics?

Description of a system

- State space
- Energy for each state
- Allowed transitions from state to state

Procedure

- Choose random initial state S_0
- Repeat as often as necessary for $i = 1, 2, 3, \ldots$:
  - Enumerate all states into which state S_{i-1} is allowed to transition.
  - Pick one of these states T randomly
  - Calculate $\Delta E \equiv E[S] - E[T]$
  - If $\Delta E \geq 0$ let $S_i \equiv T$
  - If $\Delta E < 0$ pick a random number r
  - If $r < e^{\frac{\Delta E}{RT}}$ let $S_i \equiv T$, otherwise let $S_i \equiv S_{i-1}$

Properties of Monte Carlo dynamics

Boltzmann distribution

- Transitions have to be chosen such that every state can be reached from every other state
- The sequence $S_0, S_1, \ldots$ visits each state with a frequency proportional to its Boltzmann probability $p = e^{-\frac{E[S]}{RT}}/Z$
- It avoids calculating a partition function

Monte Carlo and real time

- In general, Monte Carlo dynamics has nothing to do with real dynamics of system
- Monte Carlo dynamics is a good representation of real dynamics if:
  - The allowed transitions correspond to true elementary steps of real dynamics
  - There is a good physical reason for all downhill transitions occuring all with the same rate k_0
  - In this case, the real time is $t_{real} = i/k_0$ for the ith step

Numerical tricks

Gillespie sampling

Instead of accepting or rejecting a move, choose a move in every step but increase time by a larger chunk if moves are unfavorable.
(Gillespie, J. Phys. Chem. 1977)

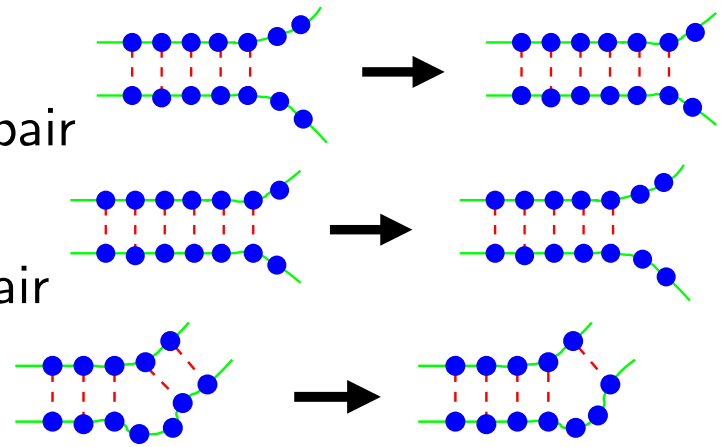
State clustering

If algorithm revisits the same state in a local minimum over and over again, precalculate the distribution and residence time in local valley by diagonalizing a reasonably sized matrix and directly choose a move that leads out of the valley.

The Vienna approach

Specification

- States: all secondary structures
- Energies: calculated by Turner model
- Transitions:
  - Opening of a base pair
  - Closing of a base pair
  - Repairing of a base



Flamm *et al.*, RNA folding at elementary step resolution, RNA 2000

kinfold program in Vienna package

Properties

Validity

- Every state can be reached from every other
- Transitions are reasonable elementary steps
- Closing of a base is an activated process itself with a roughly constant free energy barrier

Time scales

- Time for closing a base pair experimentally determined to be $\approx 1\mu s$
  $\Rightarrow$ Monte Carlo time is real time in μs
- Time for closing a hairpin of length 4:

$$e^{\frac{E[S]}{RT}}\,\mu s \approx e^{\frac{4.1}{0.6}}\,\mu s \approx 1ms$$

Isambert and Siggia approach

Specification

- Precompute all possible helices of minimum length 3
- States: all possible combinations of non-overlapping helices (Note: this includes pseudo-knots!)
- Energy:
  - Turner stacking energies for stacks
  - stick and string model for loop entropies
- Transitions: formation and removal of a complete helix

Isambert and Siggia, Modeling RNA folding paths with pseudoknots: Application to hepatitis delta virus ribozyme, PNAS 2000

Xayaphoummine *et al.*, Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots, Nucleic Acids Research 2005

Results I

Pseudoknots

- Test on 7 relatively short RNA molecules with experimentally known pseudoknots
- For 5 out of the 7 correct pseudo-knotted is found
- For 1 the pseudo-knotted structure is only marginally higher than minimum energy structure
- For 1 specific *Mg* binding is suspected

Isambert and Siggia, PNAS 2000

Results II

Kinetics

- Folding of HDV ribozyme
- 1/3 of the time folds within a fraction of a second
- 2/3 of the time trapped for up to a minute
- Molecule has to be functional after 4s to self-cleave!
- If folding co-transcriptional only 10% of molecules get trapped

Isambert and Siggia, PNAS 2000

Outline of part III

Summary of part III

- Dynamics of RNA folding is important for many biological processes
- Molecular dynamics is useful to study fast structural changes of small substructures
- Monte Carlo simulations can describe folding of realistically sized molecules
- Monte Carlo simulations can describe folding including pseudo-knots
- The elementary time scale for closing of a base pair on top of an existing step is on the order of $1\mu s$

Outline of part IV

Outline of part IV

What is a micro RNA?

Definition

A microRNA is an approximately 22 nucleotide long RNA that regulates expression of other genes

Properties

- microRNA are post-transcriptional regulators
- microRNA themselves are temporally and spatially regulated
- There are most likely hundreds of microRNA in higher eukaryotes

How are microRNA made?

A microRNA is made in four steps:

1. Transcription of the primary RNA (pri-miRNA) by RNA polymerase (in the nucleus)
2. Processing into hairpin-like RNAs called pre-miRNA (in the nucleus)
3. Cleavage into the mature microRNA by Dicer (in the cytoplasm)
4. Incorporation into the RNA-induced silencing complex (RISC)

What do microRNA do?

In plants

- microRNAs in RISC complexes bind to exactly complementary regions of the mRNAs of target genes
- The mRNA of the target gene is cut and degraded

In animals

- microRNAs in RISC complexes bind to partially complementary regions of the mRNAs of target genes
- The mRNA of the target gene is prevented from translation

Processes microRNAs are known to be involved in

- Development
- Immune system
- Cancer

Computational problems

Gene finding problem

Given a genome sequence, predict the microRNA sequences of the organism (in plants and animals).

Target prediction problem

Given a microRNA sequence, predict the genes that are regulated by that microRNA (in animals).

Here, only talk about the target prediction problem

Outline of part IV

Approaches for *Drosophila*

Many algorithms proposed essentially simultaneously

For *Drosophila*:

- Stark *et al.*, Identification of *Drosophila* microRNA targets, PLoS Biollogy (2003)
- Enright *et al.*, MicroRNA targets in *Drosophila*, Genome Biology (2003)
- Rajewsky and Socci, Computational identification of microRNA targets, Developmental Biology (2004)

Approaches for mammals

For mammals:

- Lewis *et al.*, Prediction of mammalian microRNA targets, Cell (2003)
- John *et al.*, Human MicroRNA targets, PLoS Biology (2004)
- Kiriakidou *et al.*, A combined computational-experimental approach predicts human microRNA targets, Genes & Development (2004)

All approaches rather similar, but resulting predictions have only small overlap

Here: Rajewsky and Socci

Features used in target prediction

Observations:

1. MicroRNA target sites are in 3'UTRs.
2. Although microRNAs are not exactly complementary to their targets they always contain a nucleus of length 6-8 bases with exact complementarity.
3. Even in the not exactly complementary regions, there is still a lot of stabilizing base pairing.
4. MicroRNA–target relationships tend to be conserved across species.

Nucleus identification

Using observation 2: there is a nucleus

- Assign to every base pairing stretch a score by the formula

  $$\text{score} = w_{GC}\cdot\#\text{GC base pairs} + w_{AU}\cdot\#\text{AU b.p.} + w_{GU}\cdot\#\text{GU b.p.}$$

- Use training sets of true targets and of true non-targets to choose w_{GC}, w_{AU}, and w_{GU} such that they maximize the distance between expected scores of targets and non-targets
  $\Rightarrow w_{GC} = 5,\ w_{AU} = 2,\ w_{GU} = 0$
- Fix a p-value threshold
- Score many random mRNA sequences for a given microRNA sequence and determine the score cutoff that belongs to the p-value cutoff
- Score all 3'UTR regions of interest and keep only those for which the score exceeds the threshold.

Using other features

Using observation 3: other bases pair as well

- Cut 40 base window around identified nucleus
- Use MFOLD to predict binding energy of microRNA-mRNA hybrid
- Keep only candidate targets with binding energy below -17.4kcal/mol

Using observation 4: targets are conserved

Keep only candidates for which a target is predicted in *D. melanogaster* as well as in *D. pseudoobscura*

Results

Test1

Search *D. melanogaster hid* transcript (not used in training) for *bantam* sites:

$\Rightarrow$ 4 of the 5 experimentally known sites found, no false positive

Test2

Prediction of target sites for 74 *D. melanogaster* microRNAs in 31 body patterning genes:

$\Rightarrow$ found 39 putative target sites, some of them make biological sense

Outline of part IV

Additional features used for prediction

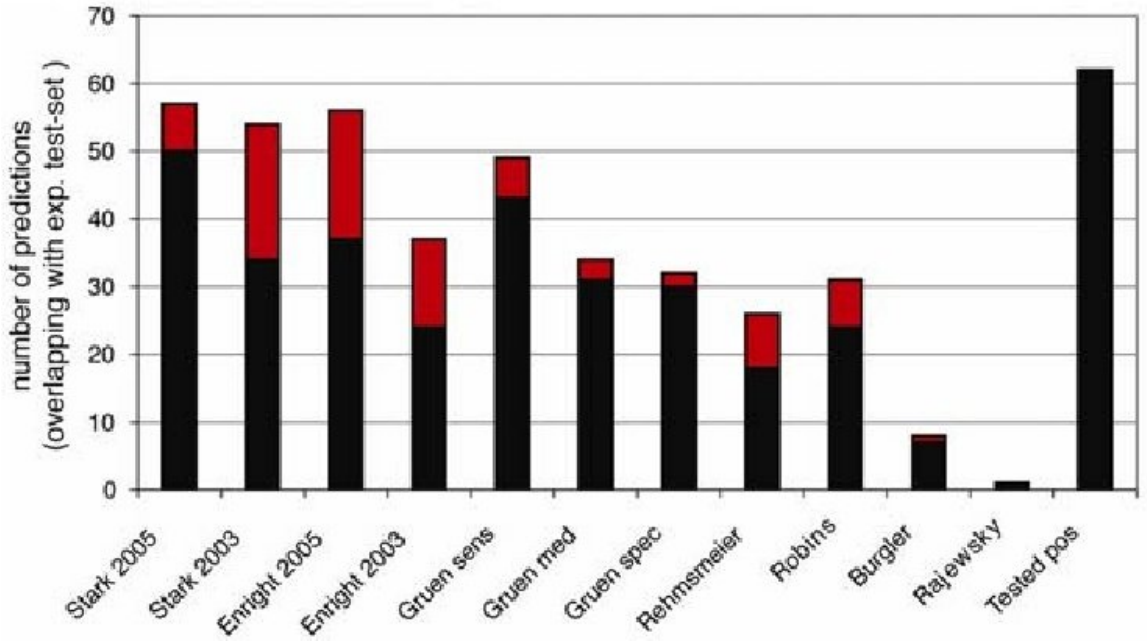
- Position of nucleus in microRNA (5' end)
- Conservation of nucleus in large multiple sequence alignments
- Absence of conservation outside nucleus
- An A immediately upstream of the nucleus
- Combinatorics of target sites for multiple microRNAs

Performance comparison

Assay

Evaluate prediction method on 133 experimentally tested targets.
(Stark, Brennecke, Bushati, Russell, Cohen, Cell 2005)

Results



Best methods:

- TargetScanS (Stark *et al.*)
- PicTar (Grün *et al.*)

Outline of part IV

Summary of part IV

- microRNAs are a relatively new but important class of regulatory RNAs
- There are many programs for microRNA target prediction
- The main features used for target prediction are stability of a nucleus, stability of the full hybrid, and conservation between species
- Best current algorithms (TargetScanS and PicTar) have about 90% accuracy and 70% – 80% sensitivity
- Reviews:
  - Issac Bentwich, Prediction of microRNAs and their targets, FEBS Letters 2005
  - Nikolaus Rajewksy, microRNA target predictions in animals, Nature Genetics 2006