Aggregation and folding phase transitions of RNA molecules

Ralf Bundschuh

The Ohio State University

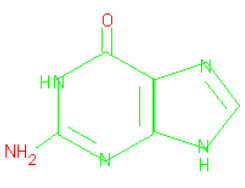
Collaborators: Vishwesha Guttal, The Ohio State University
Robijn Bruinsma, UCLA

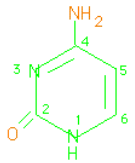
Outline:

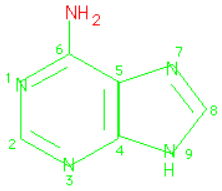
- RNA secondary structure
- Aggregation of two molecules
- Native state
- Conclusions

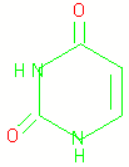
supported by the National Science Foundation

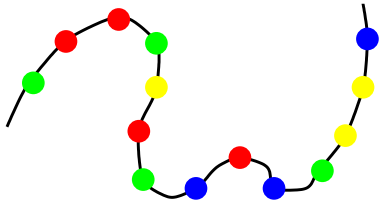
- RNA is heteropolymer of four different bases G, C, A, and U





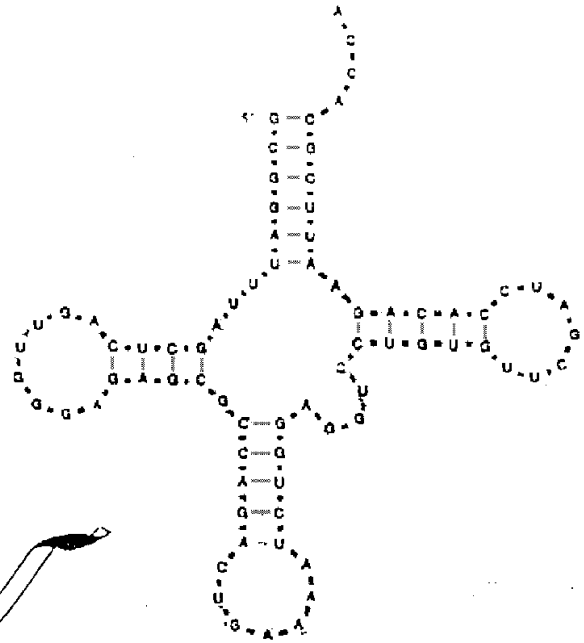






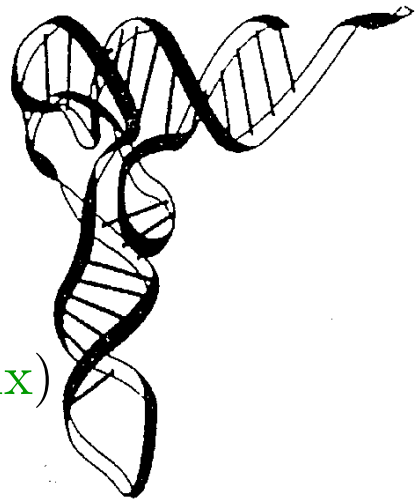
- Primary structure: Sequence, e.g.,

GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAAAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA



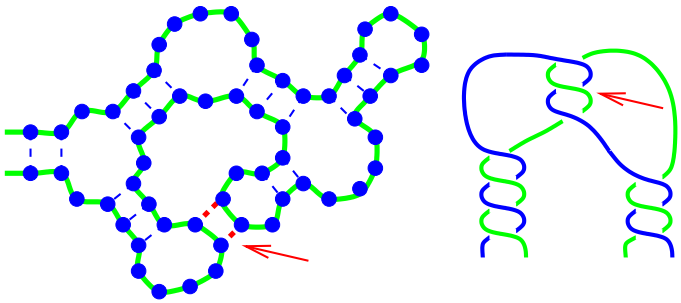
- Strongest interaction:
  Watson-Crick base pairing (G–C and A–U)
  $\longrightarrow$ secondary structure

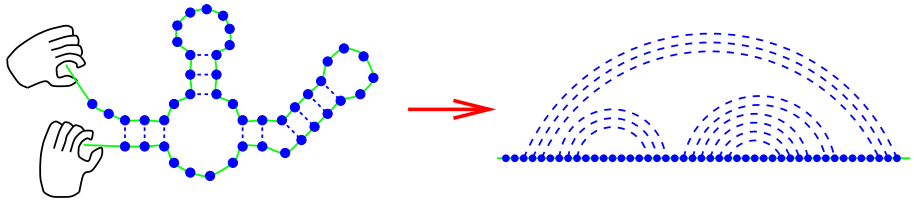
- Spatial arrangement
  $\longrightarrow$ tertiary structure
  (looks locally like DNA double helix)



Introduction II

- Secondary structure: Set of base pairs formed
- Pseudo-knots neglected
- Diagramatic representation

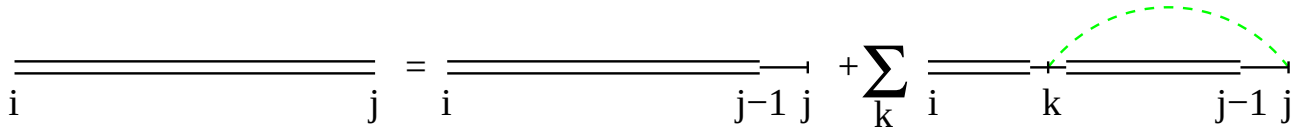




- Assign energy $E[S]$ to each structure S $\longrightarrow$partition function

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

- Partition function generated exactly by Hartree equation



$\rightarrow$ Electron in disordered medium

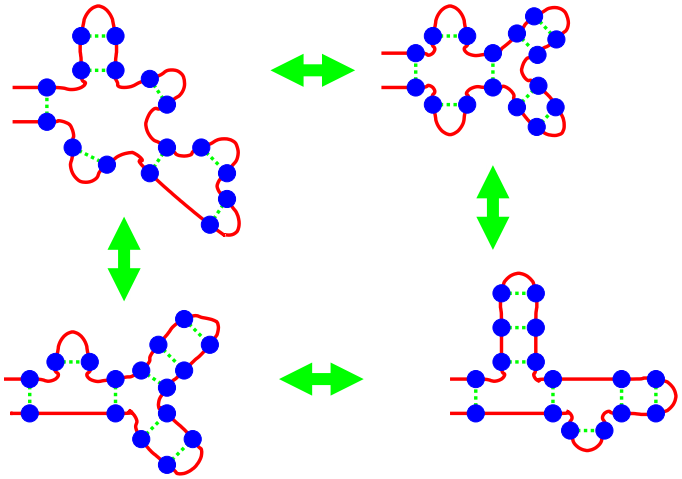
- Handle for analytical treatment
- $O(N^3)$ algorithm for exact partition function (McCaskill, Biopolymers 29, 1990.)

- Many biological functions: Making proteins, structural RNA, gene regulation, ...
- Also interesting from the statistical physics point of view:
  - Entropy (many different structures)
  - Energy (systematic preference(s) for certain structures)
  - Disorder (through sequences)

$\Rightarrow$ Phases, phase transitions, critical exponents in the limit $N \to \infty$

- Here: no disorder
- Assume uniform interaction between any two bases
- Justified for uniform sequences AUAUAUAUAU... or GCGCGCGCGC...
- Also believed to be good description of random sequences at high temperatures on a coarse grained scale

- Quantification of the molten phase
- Uniform attraction between any two elements of the polymer
  $\longrightarrow$ only one interaction parameter ε_0
  $\longrightarrow$ or one Boltzmann factor $q \equiv e^{-\varepsilon_0/T}$
- Main effect in molten phase: branching entropy



- Hartree equation  becomes

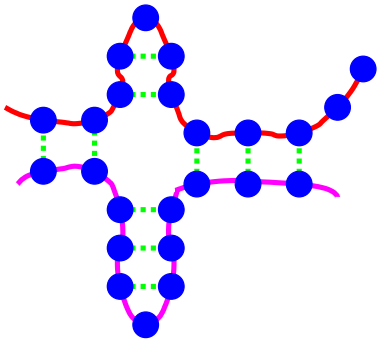
$$G_0(N+1;q) = G_0(N;q) + q\sum_{k=1}^{N} G_0(k;q)G_0(N-k;q)$$

$$\Rightarrow \quad \widehat{G}_0(z;q) = z\widehat{G}_0(z;q) - 1 - q\widehat{G}_0(z;q)^2 \quad \text{in } z \text{ domain}$$

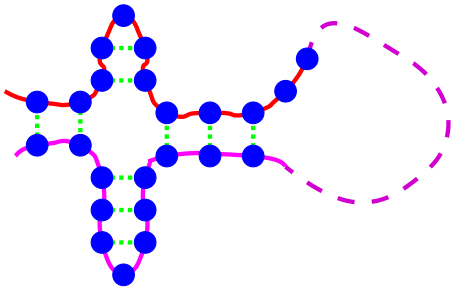
$$\Rightarrow \quad G_0(N;q) \sim N^{-\theta} z_0^N \text{ with } \theta = \frac{3}{2}$$

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

- Aggregation of two RNA molecules
- Riboswitches, genes involved in Huntington's disease, model for prion diseases
- Three different interactions:
  - Within RNA molecule 1: Boltzmann factor q_1
  - Within RNA molecule 2: Boltzmann factor q_2
  - One base from each RNA molecule: Boltzmann factor q_3



- Merge the two molecules at the end
  $\longrightarrow$ can calculate $Z(N; q_1, q_2, q_3)$
  by same Hartree equation



- How does $Z(N; q_1, q_2, q_3)$ behave for large N?

- Limiting cases:

- $q_1 = q_2 = q_3 \equiv q$:
- Same as a single RNA in molten phase

$$\Rightarrow \quad Z(N; q, q, q) = G_0(2N; q) \sim N^{-3/2} {z_0}^{2N} \quad \Rightarrow \quad \theta = \frac{3}{2}$$

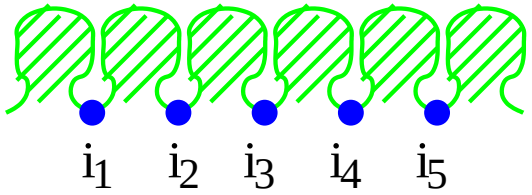
- $q_3 = 0$:
- Two non-interacting RNA molecules

$$\Rightarrow \quad Z(N; q_1, q_2, 0) = G_0(N; q_1) G_0(N; q_2) \sim N^{-3} z_1^{2N} \quad \Rightarrow \quad \theta = 3$$

- General case: fix number k of intermolecular base pairs
- Each molecule can independently explore the possible structures with any configuration of the k bases in contact with the other molecule:

$$Z_1(N;k,q) \equiv \sum_{i_1<i_2<\ldots<i_k} G_0(i_1-1;q)G_0(i_2-i_1-1;q)\ldots G_0(N-i_k;q)$$

- Total partition function by summing over all k:

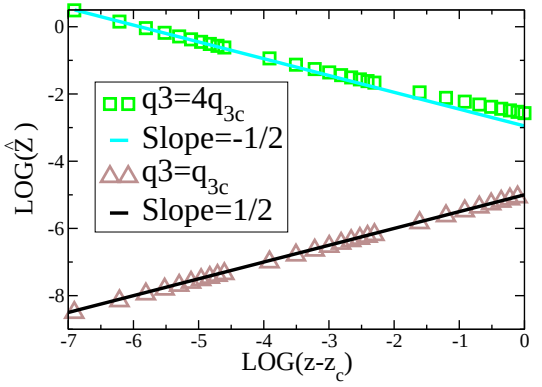


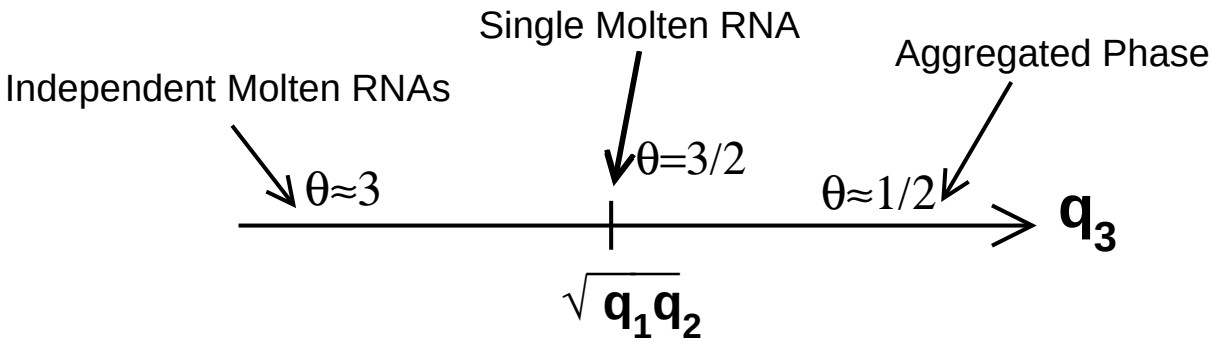
$$Z(N;q_1,q_2,q_3) = \sum_{k=0}^{\infty} {q_3}^k Z_1(N;k,q_1)Z_1(N;k,q_2)$$

- In z-space:

$$\begin{aligned}
\widehat{Z}(z;q_1,q_2,q_3) &= \sum_{k=0}^{\infty} {q_3}^k \oint \frac{\mathrm{d}z'}{z'} \widehat{Z_1}(z;k,q_1)\widehat{Z_1}(z/z';k,q_2) \\
&= \sum_{k=0}^{\infty} {q_3}^k \oint \frac{\mathrm{d}z'}{z'} \widehat{G_0}(z;q_1)^{k+1}\widehat{G_0}(z/z';q_2)^{k+1} \\
&= \oint \frac{\mathrm{d}z'}{z'} \frac{\widehat{G_0}(z;q_1)\widehat{G_0}(z/z';q_2)}{1-q_3\widehat{G_0}(z;q_1)\widehat{G_0}(z/z';q_2)}
\end{aligned}$$

- Evaluate complex contour integral numerically
- Result: phase transition at $q_{3c} \equiv \sqrt{q_1 q_2}$
  - $\theta = 3$ for $q_3 < q_{3c}$ (two independent RNA molecules)
  - $\theta = 3/2$ for $q_3 = q_{3c}$ (behavior of one RNA molecule)
  - $\theta = 1/2$ for $q_3 > q_{3c}$ (new aggregated phase)

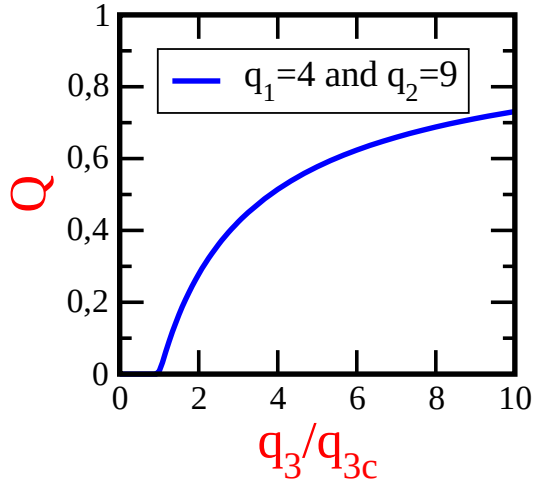


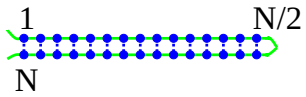


- Order parameter: fraction of intermolecular base pairs

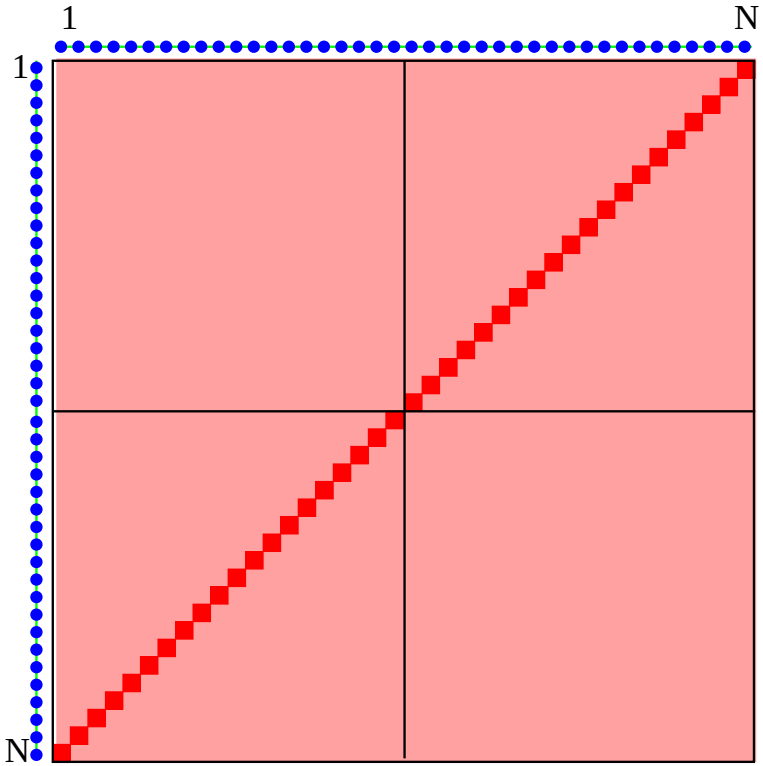
$$Q \equiv - \lim_{N \to \infty} \frac{1}{N} \frac{\partial \ln Z(N; q_1, q_2, q_3)}{\partial \ln q_3}$$

$\Rightarrow$ Second order phase transition



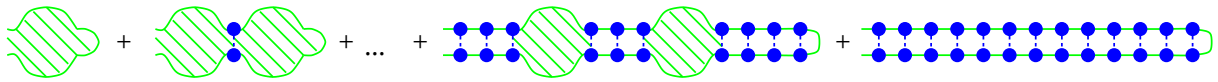
- How does native structure appear ?
- Model bias towards native structure
- Case 1: Native structure is long hairpin 

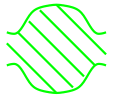
- Two Boltzmann factors
  - Native base pair: Boltzmann factor $\tilde{q}$
  - Generic base pair: Boltzmann factor q



- Model similar to Gō model of protein folding (Gō, J. Stat. Phys. 30, 1983)

- Review of RB and Hwa, PRL 1999:
- Partition function: order arbitrary structure by number of native contacts

$$Z(N; q, \tilde{q}) =$$ 

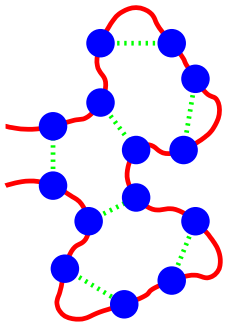
-  $= W(\ell) =$ sum over all ways to place non-native bonds $\approx G_0(2\ell; q)$
- Relation between bubble (W) and full (Z) partition functions

$$\widehat{Z}(z; q, \tilde{q}) = \widehat{W}(z) + \widehat{W}(z)\,\tilde{q}\,\widehat{W}(z) + \widehat{W}(z)\,\tilde{q}\,\widehat{W}(z)\,\tilde{q}\,\widehat{W}(z) + \ldots$$

$$= \frac{\widehat{W}(z)}{1 - \tilde{q}\widehat{W}(z)}$$

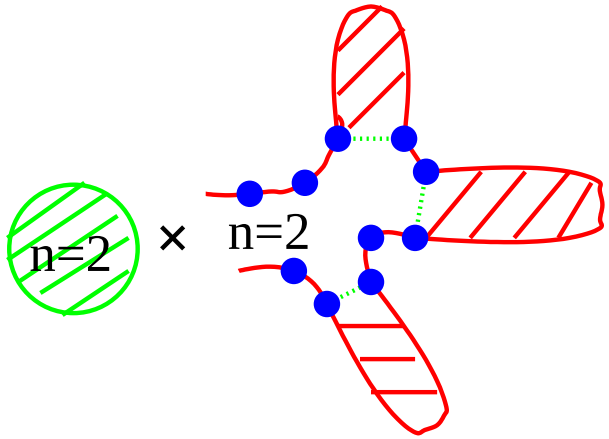
- Exact expression for $\widehat{Z}(z; q, \tilde{q})$
- Phase transition between molten and native phase at finite critical bias $\tilde{q}_c$ or at critical temperature T_c
- Phase transition is second order with finite jump in specific heat ($\alpha = 0$)

- But: Natural RNA structures are branched



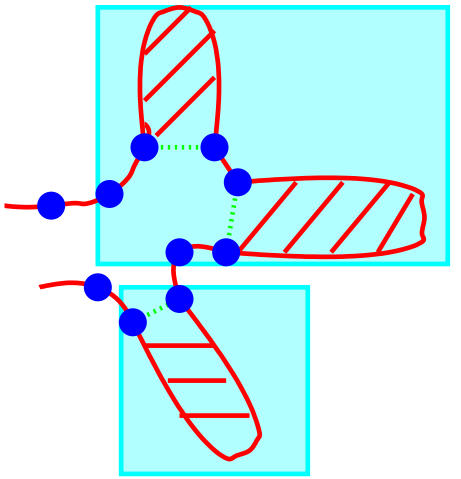
$\Rightarrow$ Case 2: Cayley tree as native structure

- Characterize by order k instead of length $N = 2^{k+2} - 2$
- Define $G(k, n)$ as the partition function of all structures with n bases "outside" the native base pairs and all these bases unpaired
- $Z(k; q, \tilde{q}) = \sum_n W(n) G(k, n)$

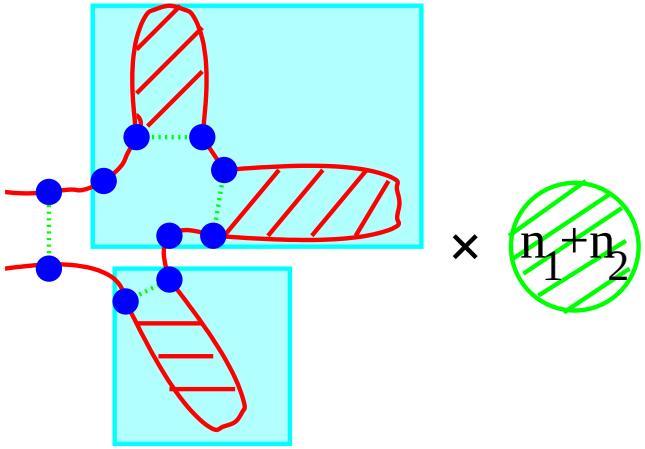


Native state IV

- Recursion equations from level k to $k+1$:
  - For $n > 0$ the first native base pair is open
    $\longrightarrow$ both sub-trees contribute to accessible base pairs
    $\longrightarrow G(k+1,n) = \sum_m G(k,m)G(k,n-m)$



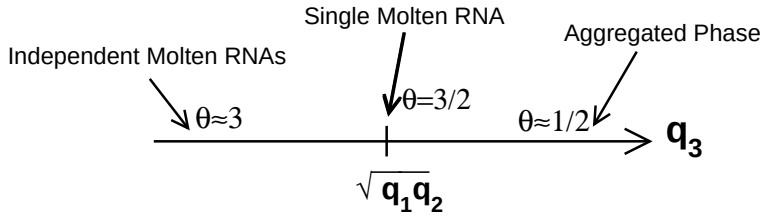
  - For $n = 0$ the first native base pair is paired
    $\longrightarrow$ both sub-trees contribute to
    first bubble of non-native base pairs
    $\longrightarrow G(k+1,n=0) =$
    $\sum_{n_1,n_2} G(k,n_1)G(k,n_2)W(n_1+n_2)$



- Convolutions $\longrightarrow$ z-transform $\longrightarrow$ self-consistency equation $\longrightarrow$ numerically solvable
- Phase transition between molten and native phase at finite critical bias $\tilde{q}_c$ or at critical temperature T_c
- Phase transition is most likely first order

- Summary:
  - RNA secondary structure formation is tractable analytically and numerically by methods of statistical mechanics

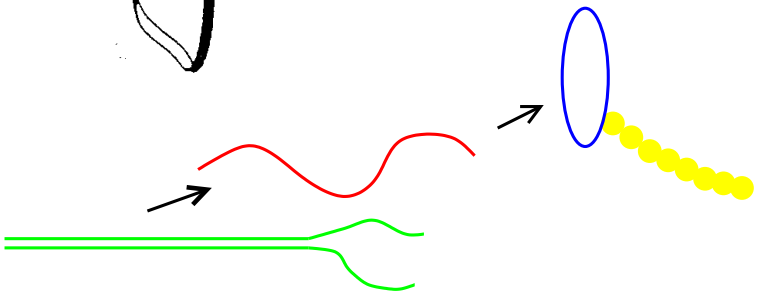
  - RNA secondary structure formation shows very rich phase behavior



- Future work:
  - Combining native states and aggregation
  - disorder
  - kinetics

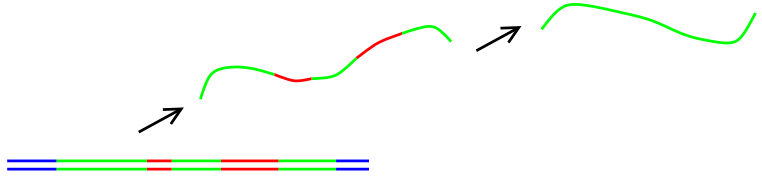
- Biological functions:

- Structure ($\longrightarrow$ proteins)
  - Ribosomal RNA
  - Transfer RNA

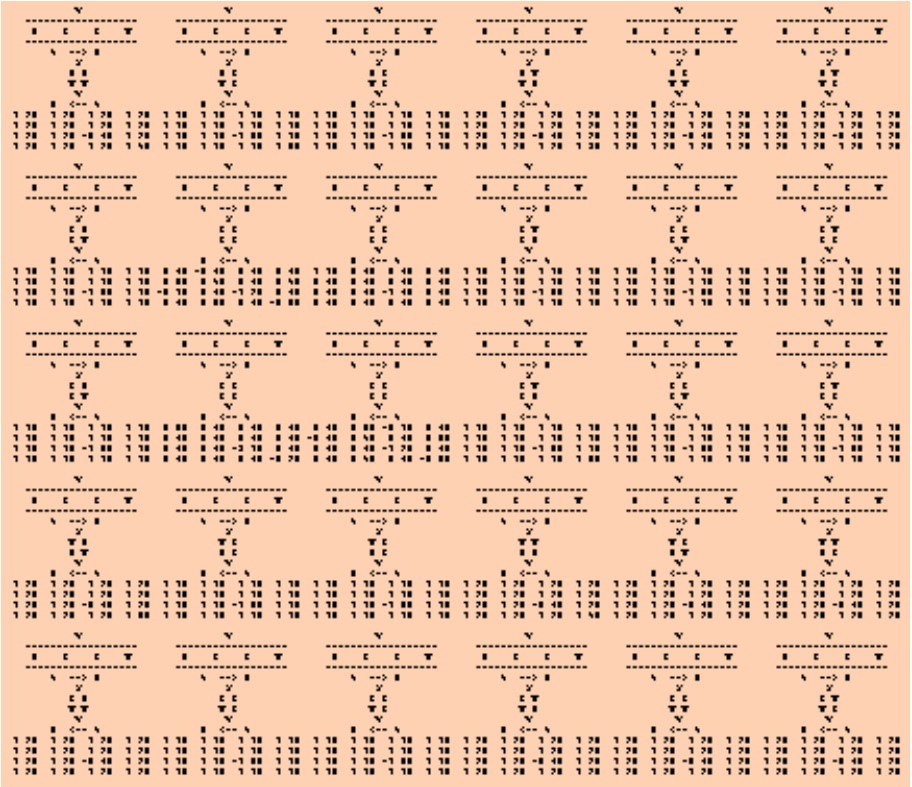


- Information ($\longrightarrow$ DNA)
  - Messenger RNA
  - single-stranded DNA
    - T instead of U
    - more rigid backbone

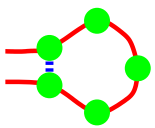
- Interplay of structure and information
  - Splicing
  - Ribozymes
  - RNA world (origin of life)



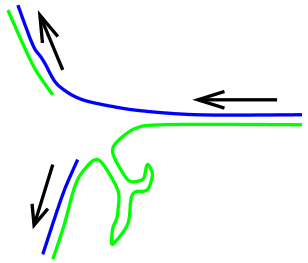
- Application to real sequences
- Use experimentally determined parameters from RNA secondary structure prediction which take all energetic details into account



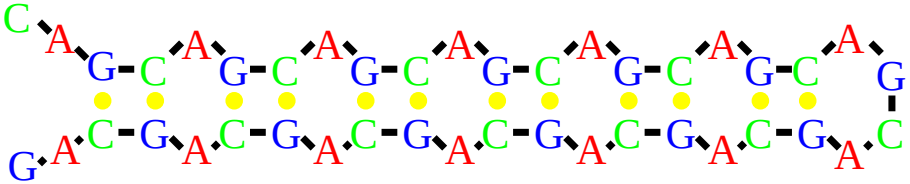
Hofacker et al., Monatshefte f. Chemie 125, 1994

- Uniform sequences AUAUAUAUAU... and GCGCGCGCGC... need very long sequences ($\geq$ 8000 bases, Tsunglin Liu & RB, PRE 2005)
  - Hairpin loops must contain at least 3 bases 
  - Loss in binding energy large in hairpin loops

- Naturally occuring in human DNA: $(\mathtt{CAG})_n$ 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 $\longrightarrow$ 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



- 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