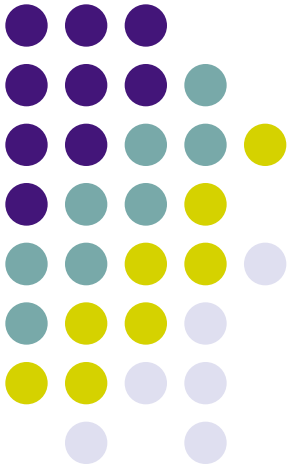
COMP598: Advanced Computational Biology Methods & Research

Introduction to RNA secondary structure prediction



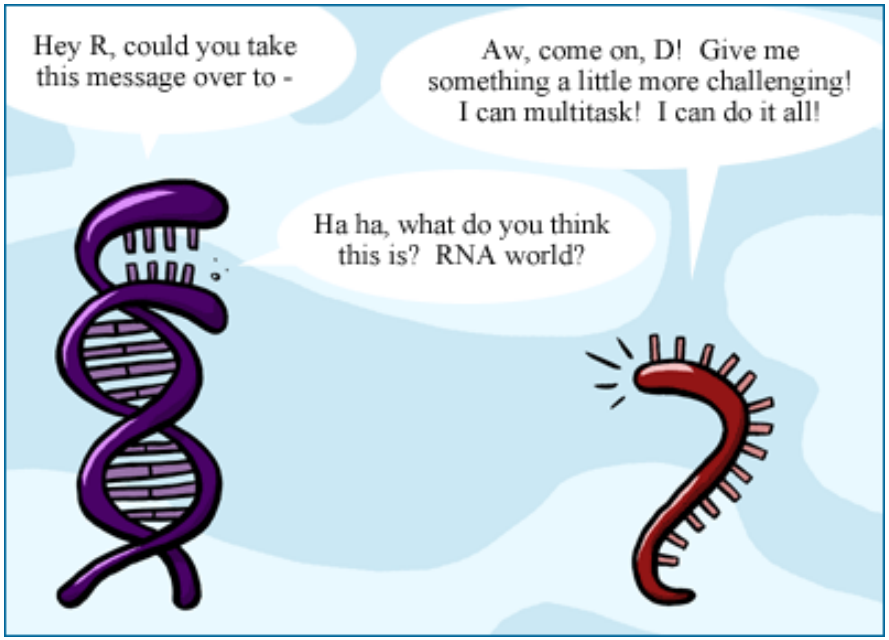
Jérôme Waldispühl
School of Computer Science, McGill

RNA world

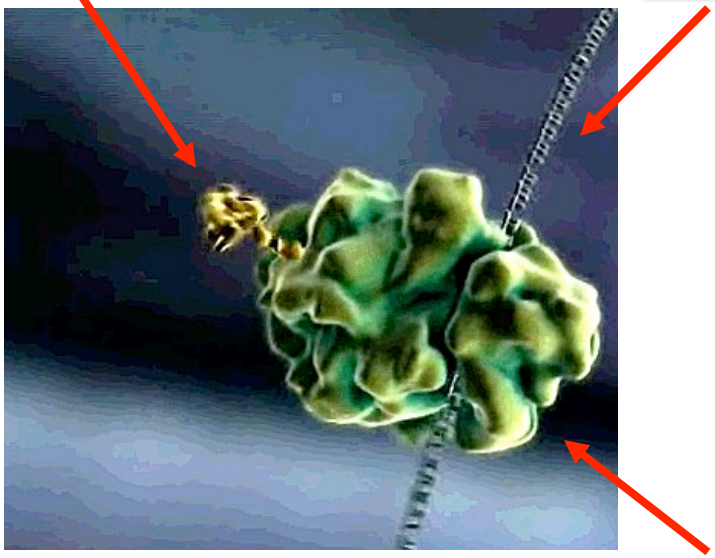
In prebiotic world, RNA thought to have filled two distinct roles:

1. an information carrying role because of RNA's ability (in principle) to self-replicate,
2. a catalytic role, because of RNA's ability to form complicated 3D shapes.

Over time, DNA replaced RNA in
Its first role, while proteins replaced
RNA in its second role.

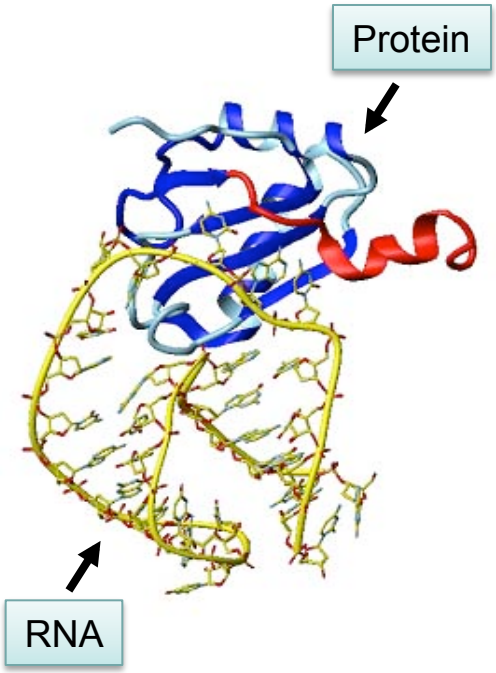


RNA classification



Messenger RNA:

- Carry genetic information,
- Structure less important.



Non-coding RNA:

- Functional,
- Structure is important.

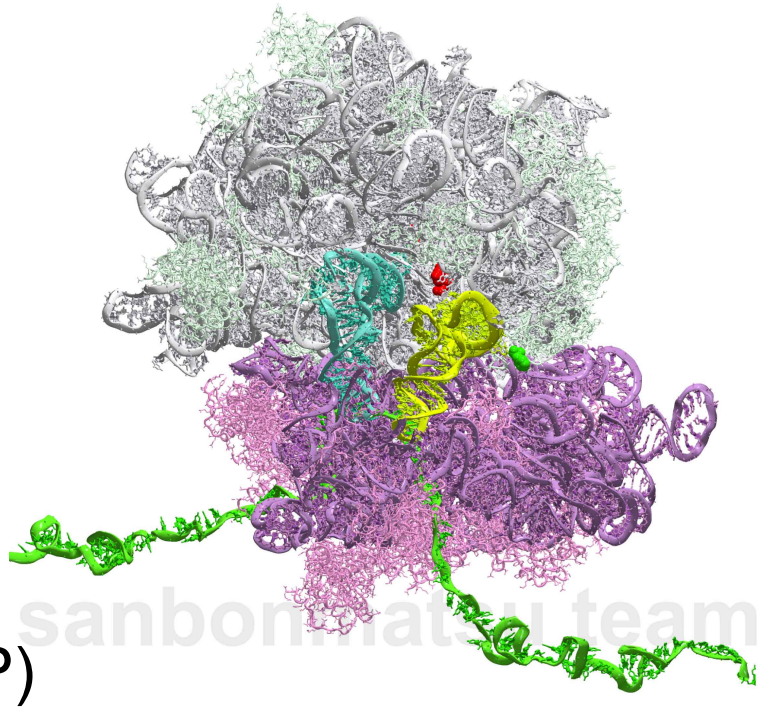
Cellular functions of RNA

Genetic Functions:

- Messenger RNA
- Viroids
- Transfer RNA

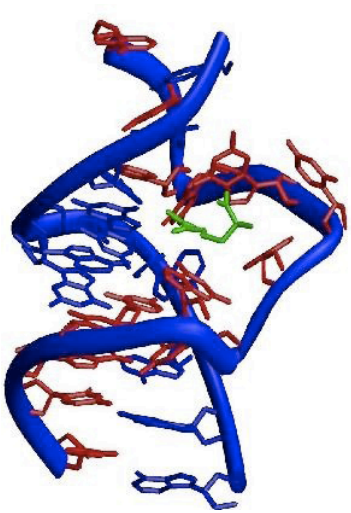
Enzymatic functions:

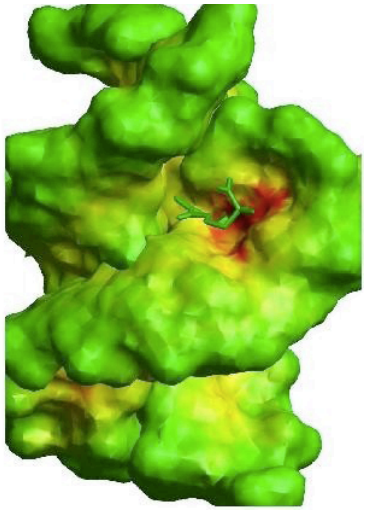
- Splicing (snRNA)
- RNA Maturation (ribonuclease P)
- Ribosomic RNA
- Guide RNA (snoRNA)

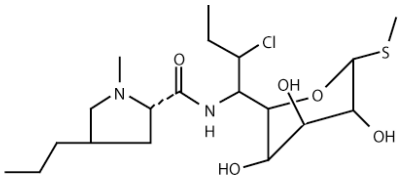


RNA structure and function

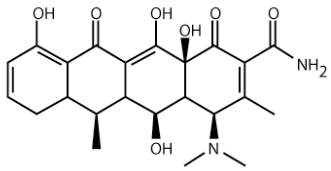
- RNAs have a 3D structure,
- This 3D structure allow complex functions,
- The variety of RNA structures allow the specific recognition of a wide range of ligands,
- Some molecules target these RNA structures (antibiotics, antimitotics, antiviruses):



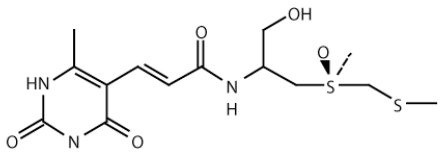




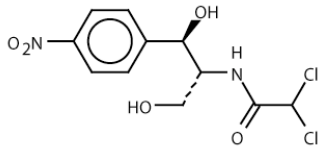
Linezolid



Doxycyclin

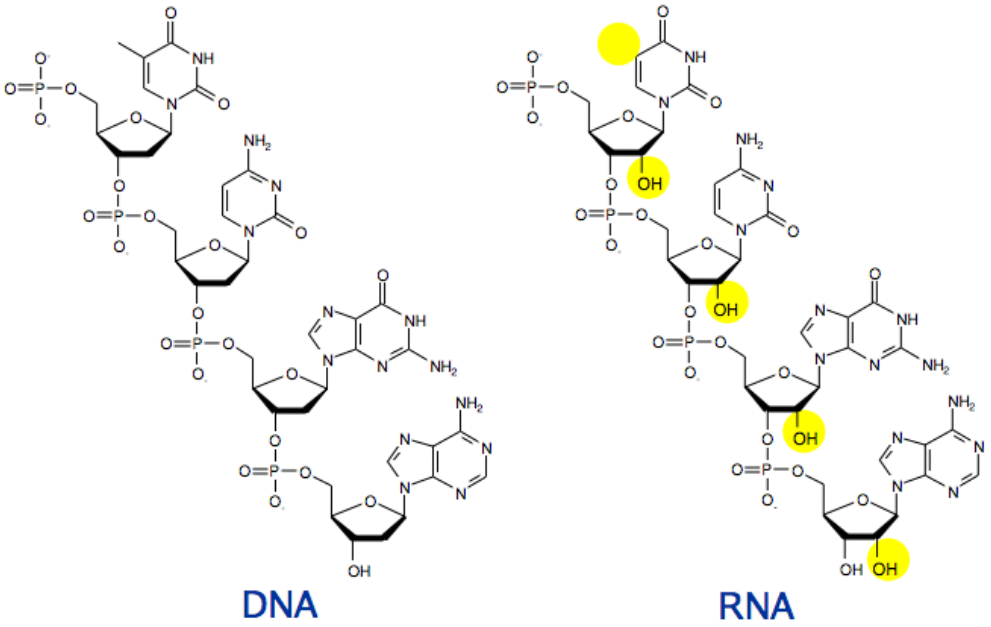


Sparsomycin



Chloramphenicol

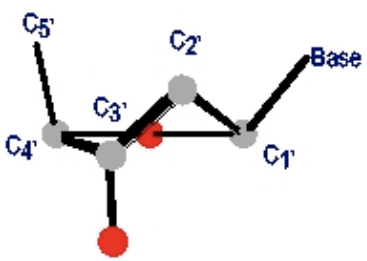
RNA vs DNA: Chemical nature



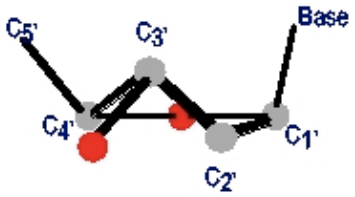
- 2’-OH group attached to sugar (instead of 2’-H): *more polar*
- Substitution of thymine by uracile = suppression of group 5-CH3

Small modifications => big effects

RNA vs DNA: Modification of the local and global geometry



C2' endo



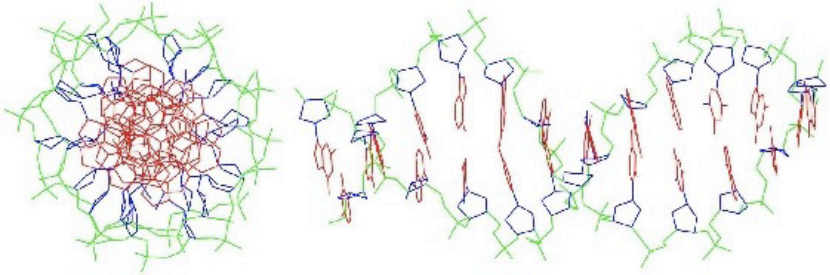
C3' endo

:Local conformation

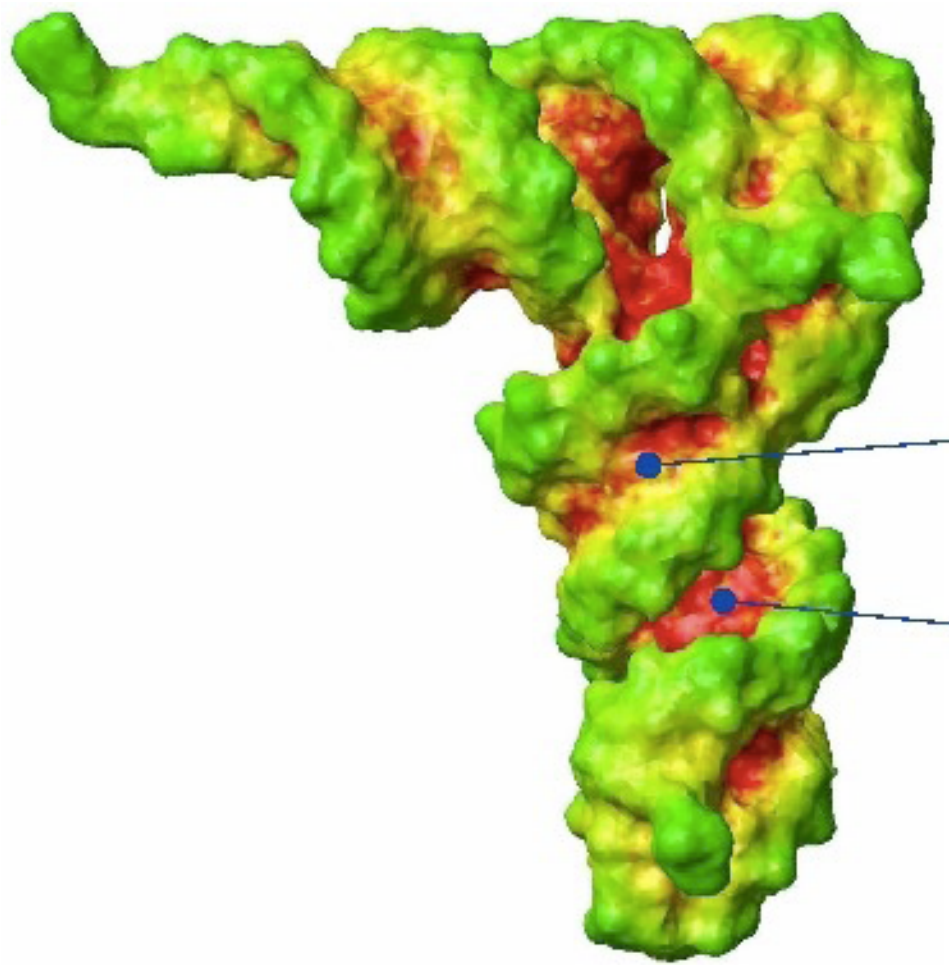
Global conformation:

RNA favorite:

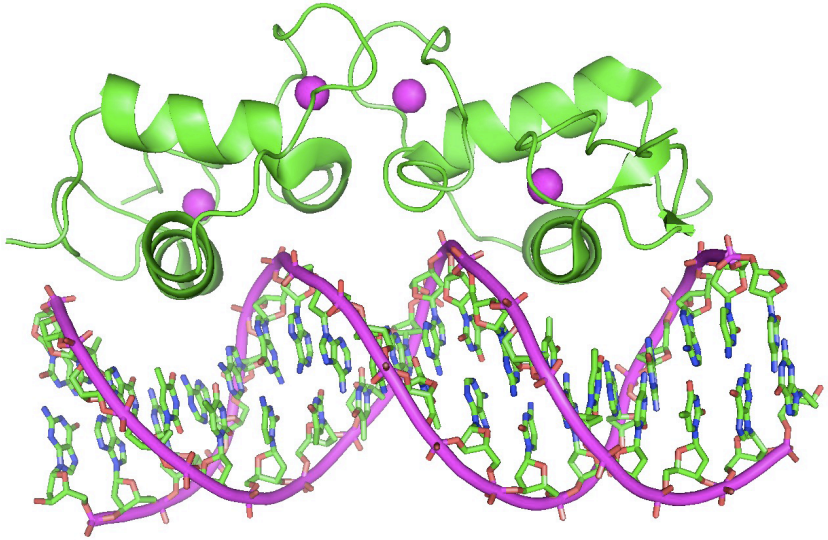
DNA favorite:



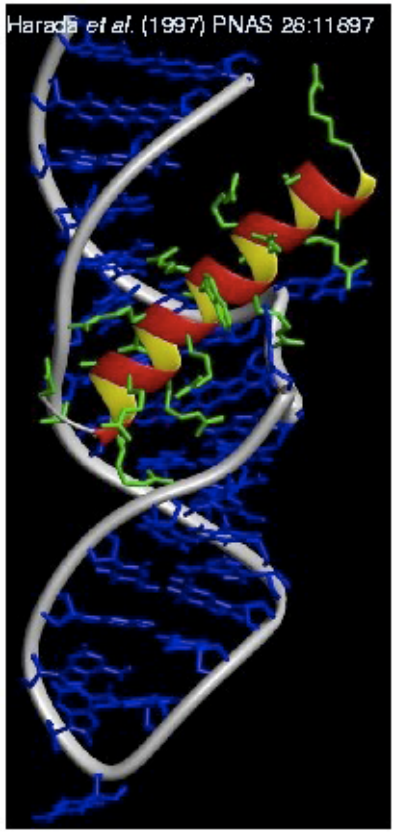
RNA vs DNA: Consequence of the modification of the geometry



RNA vs DNA: RNA-Protein and DNA-Protein interactions are different



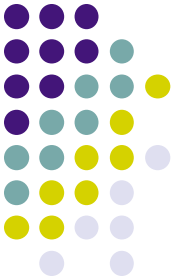
DNA-Protein: Secondary structure elements insert in big furrow



Protein binds to an irregularity of the helix

RNA-Protein interaction are more specific. Usually using less structured regions.

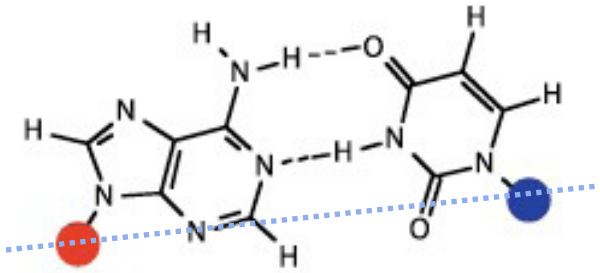
RNA vs DNA: Last (?) differences



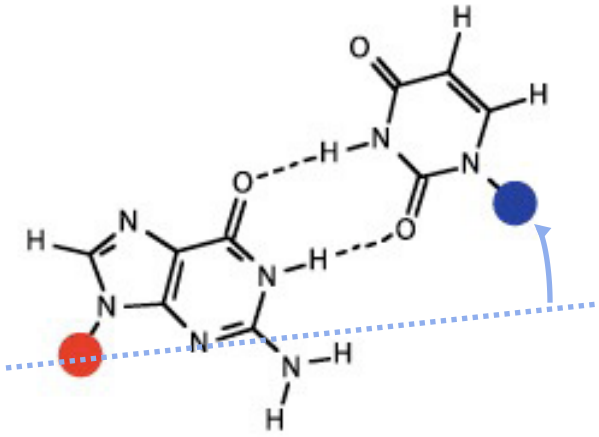
- RNA is a short linear molecule
  DNA long ≠ RNA short
- RNA are usually single stranded
  ADN double stranded ≠ ARN single stranded
- « turnover » relatively fast
  ADN stable ≠ ARN versatile

Base pairing in RNAs

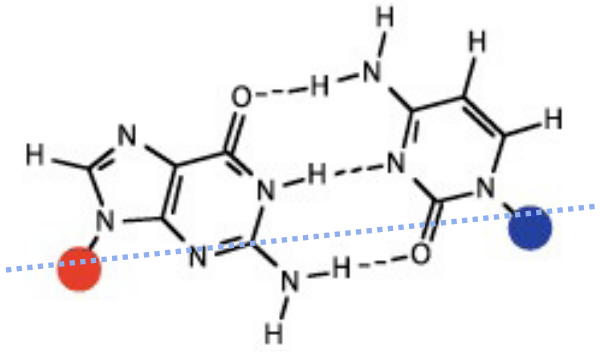
- As in DNA, bases can interact through hydrogen bonds.
- Beside the two canonical base-pairs, RNA structure allows “Wooble” base-pairs.
- A-U and G-C are “isosterus” while G-U induce a distortion of the backbone.



Paire A-U

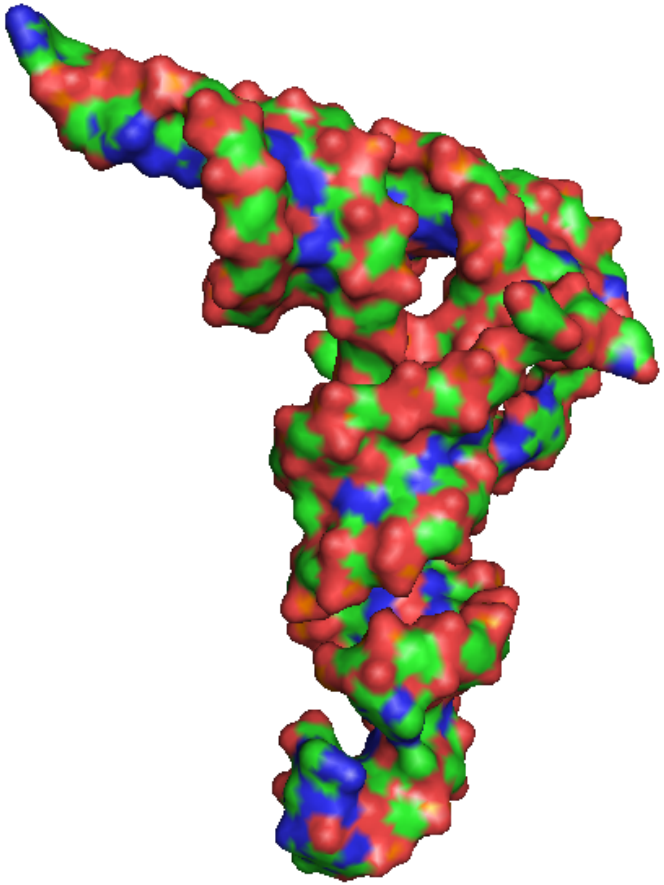


Paire "bancale" G-U

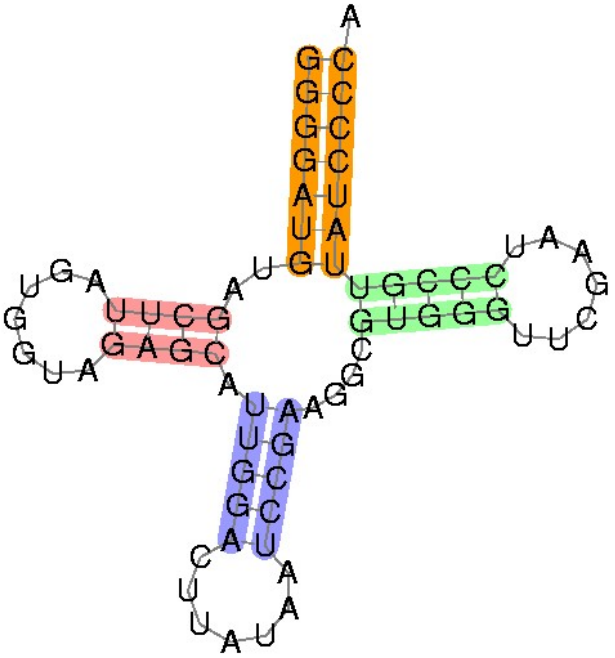


Paire G-C

RNA secondary structure



The **secondary structure** is the ensemble of base-pairs of the structure.



RNA secondary structure

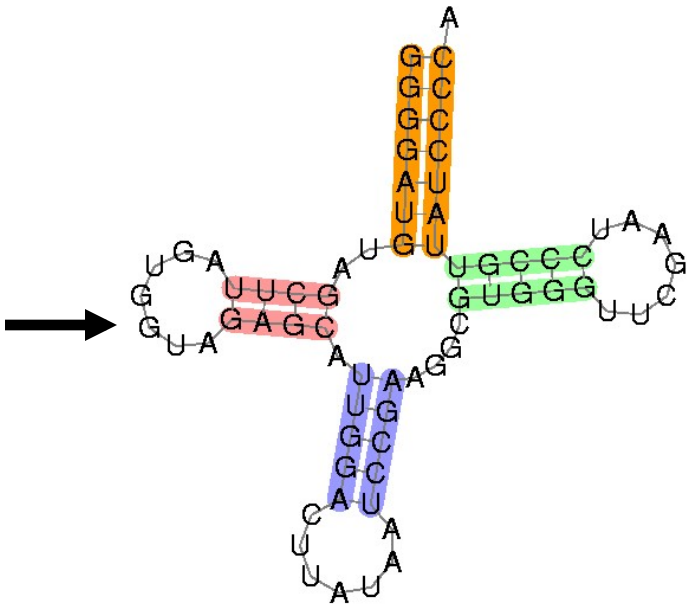
Central assumption: RNA secondary structure forms before the tertiary structure.

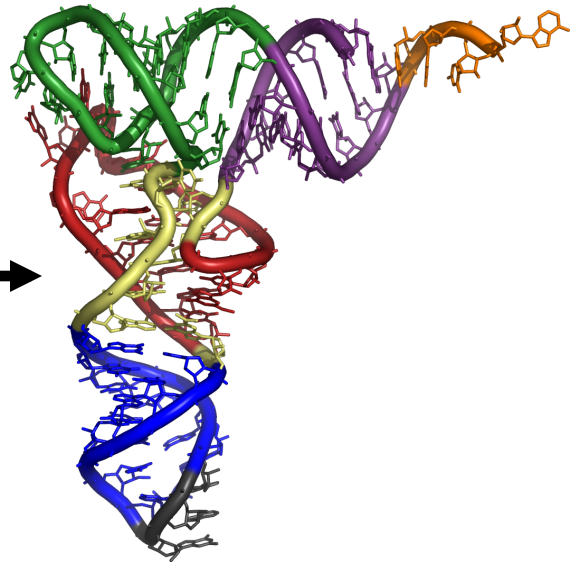
Primary structure

Secondary structure

Tertiary structure

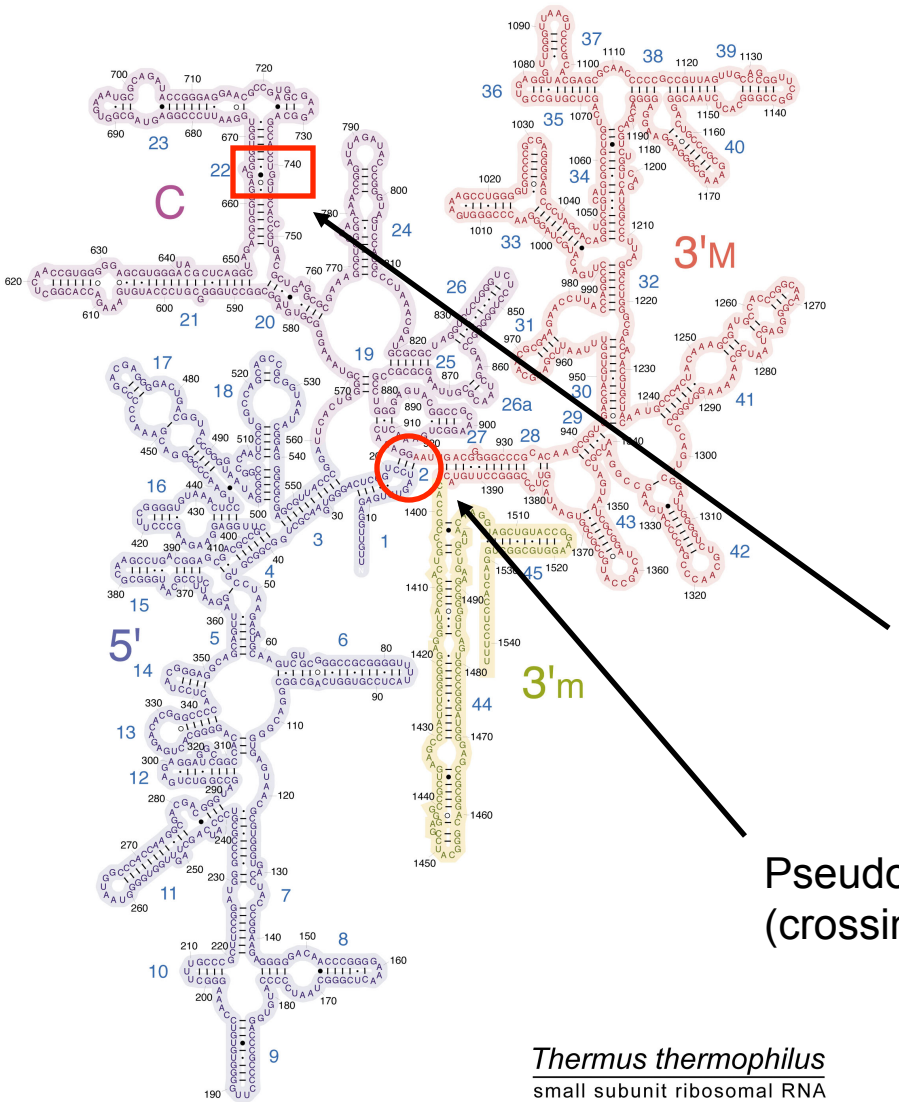
cgcgggggttgatataatataaaaaataat
aaataataataataataattatcatcatt
tccgacccatattataataatacgggttg
gaaatatagatataatatttattatattga
tataatacatatatataagttagaggaaa
tgttgtttaaaggttaaactgttagattgc
aaatctacacatttagagttcgattctctt
catttcttatatatatactacccacgcg





Secondary structure prediction is an important step toward 3D structure prediction.

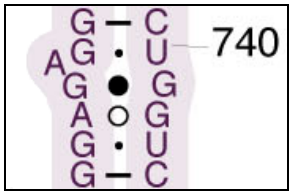
RNA secondary structure



Thermus thermophilus
small subunit ribosomal RNA

The secondary structure can be very complex. Usually most of it can be drawn on a plane.

Few “irregularities” remain.

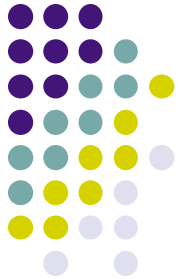


Non-canonical base-pairs

Pseudo-knot
(crossing interaction)

Base triplets
(Not on the picture)

Pseudo-knot free RNA secondary structure



Assumption: The "backbone" of the RNA secondary structure does not contain pseudo-knots, triplets and non-canonical base pairs.
(to be discussed later…)

Definition [Secondary structure without pseudo-knot]:
The secondary structure *without pseudo-knot* of an RNA sequence $a_1 \ldots a_n \in \{A,C,G,\}^n$ is an undirected graph $G = (V;E)$, where $V = \{1, \ldots, n\}$, $E \subseteq V \times V$, such that:

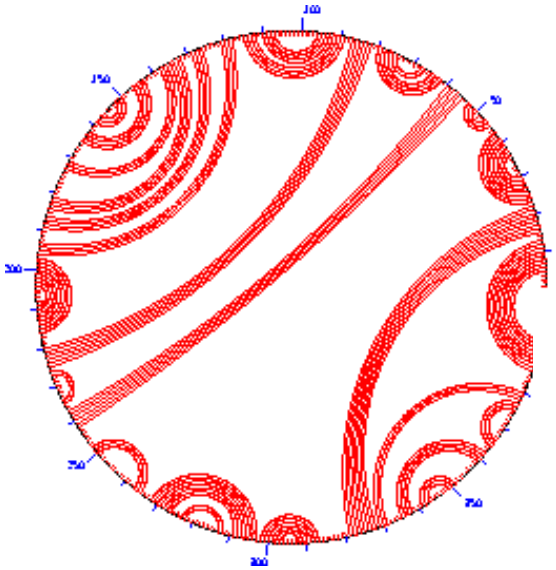
1. $(i,j) \in E \Leftrightarrow (j,i) \in E$.

2. $\forall\ 1 \leq i < n, (i; i + 1) \in E$.

3. For $1 \leq i < n$, there exists at most one $j \neq i \pm 1$ for which $(i,j) \in E$ (no triplets, etc.).

4. If $1 \leq i < k < j \leq n$, $(i,j) \in E$ and $(k,l) \in E$, then $i \leq l \leq j$
   (no knots or pseudo-knots).

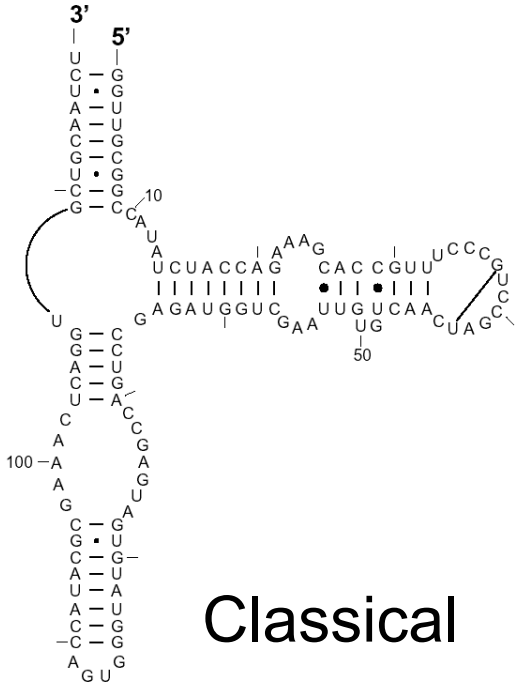
RNA secondary structure representations



Circular

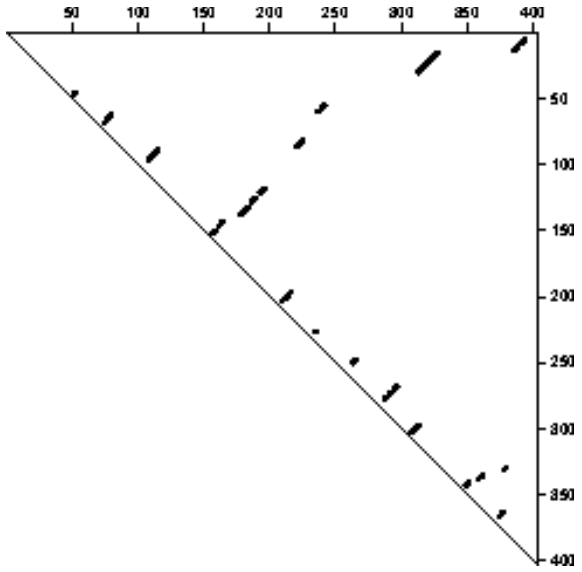
Brackets

..((((((((.(((..((...)))))...(((....))))).)))))



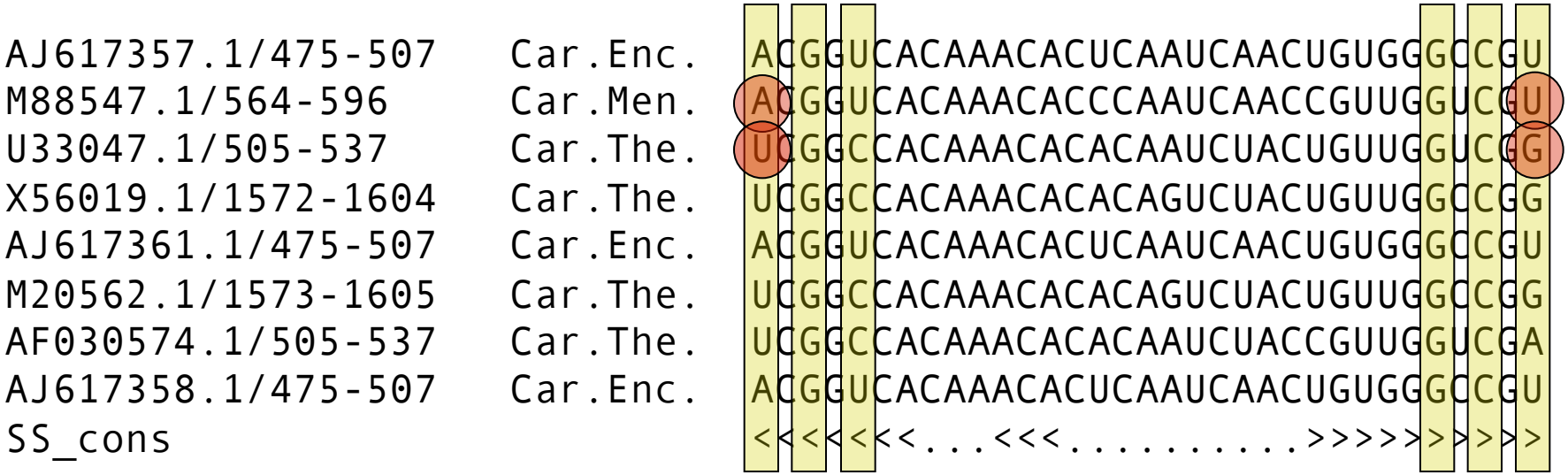
Classical

Dot plot



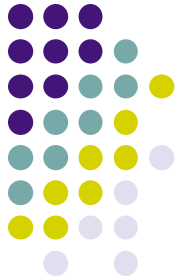
RNA secondary structure prediction using comparative methods

The secondary structure can be predicted from the alignment of homologous sequences. Base-pairs are identified through compensatory mutations.



97% of the base pairs predicted by comparative analysis in rRNAs have been confirmed later in the crystal structure.

RNA secondary structure Prediction: Part I



Aim 1: Compute the secondary structure with the maximal number of canonical base pairs (Nussinov-Jacobson, 1980).

Algorithm (Nussinov-Jacobson):

- $M_{i,j}=0$ if $j \leq i+1$,
- $M_{i,j}= \max(M_{i,j-1}, \max_{i \leq k < j}(1+M_{i,k-1}+M_{k+1,j-1}$, *if (k,j) base pair*).

j does not base pair.

j base pair between i and j-1.

RNA secondary structure prediction: Part I

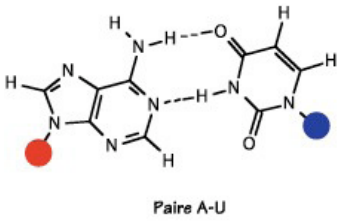
Proof: Exercise!!

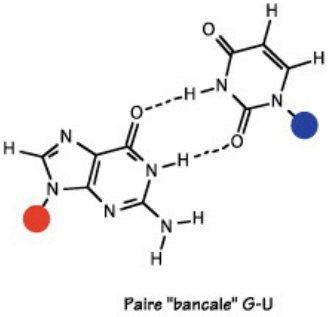
Limitations: Accuracy is low.

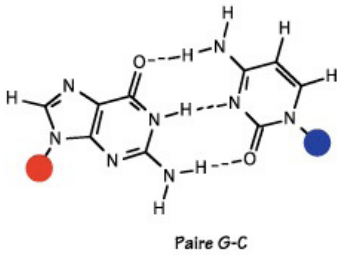
Improvements: Weight the base pairs differently.

(G-C) and (C-G): 3
(A-U) and (U-A) : 2
(G-U) and (U-G): 1

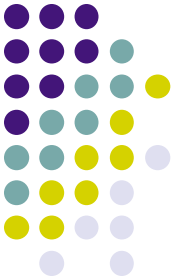
(Number of h-bonds in the base pair)







RNA nearest neighbor energy model



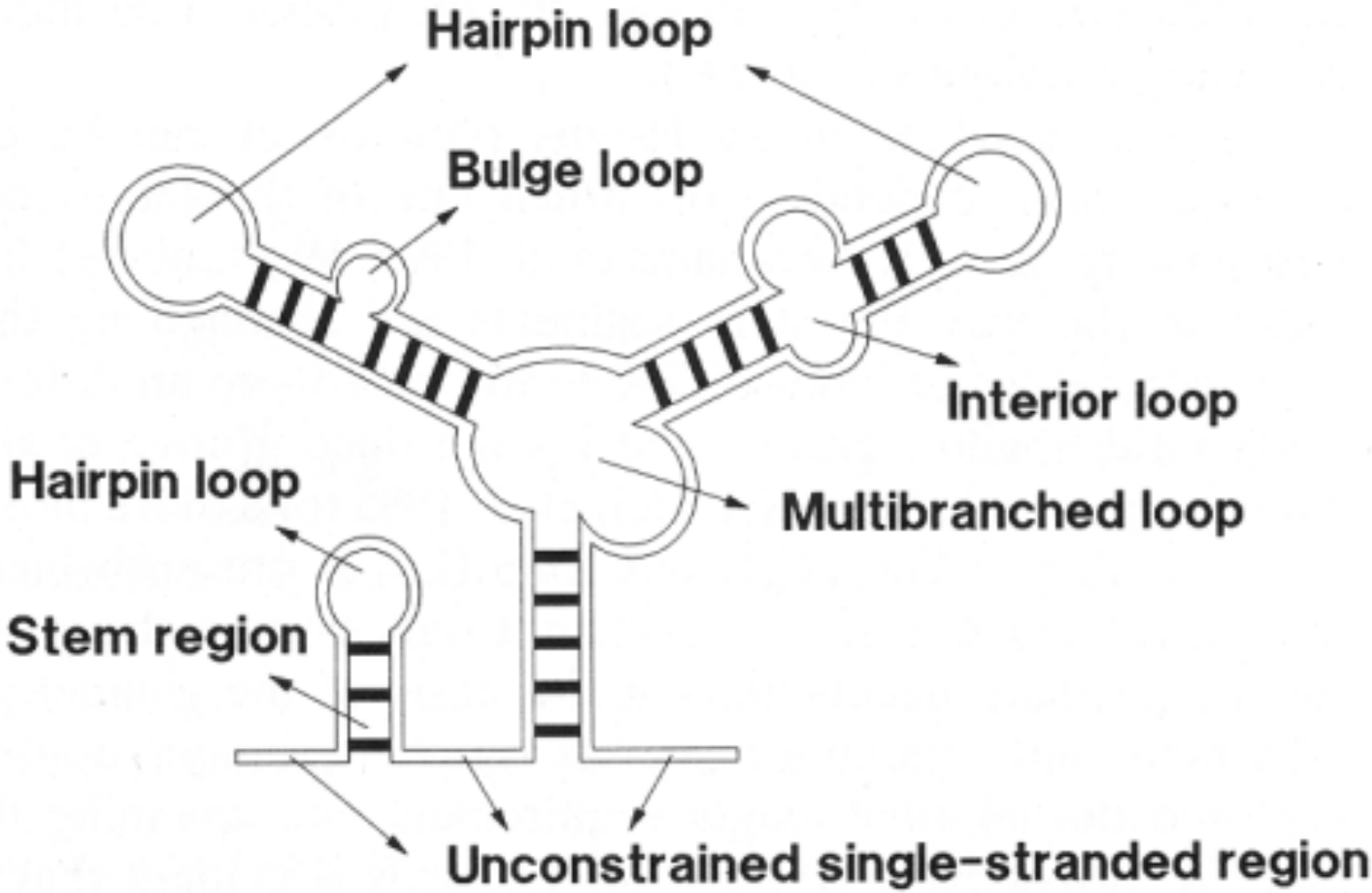
But the accuracy is still moderate. We need a better model to weight the structures.

How?: Derive a thermodynamical energy model from experimental measures (Turner group).

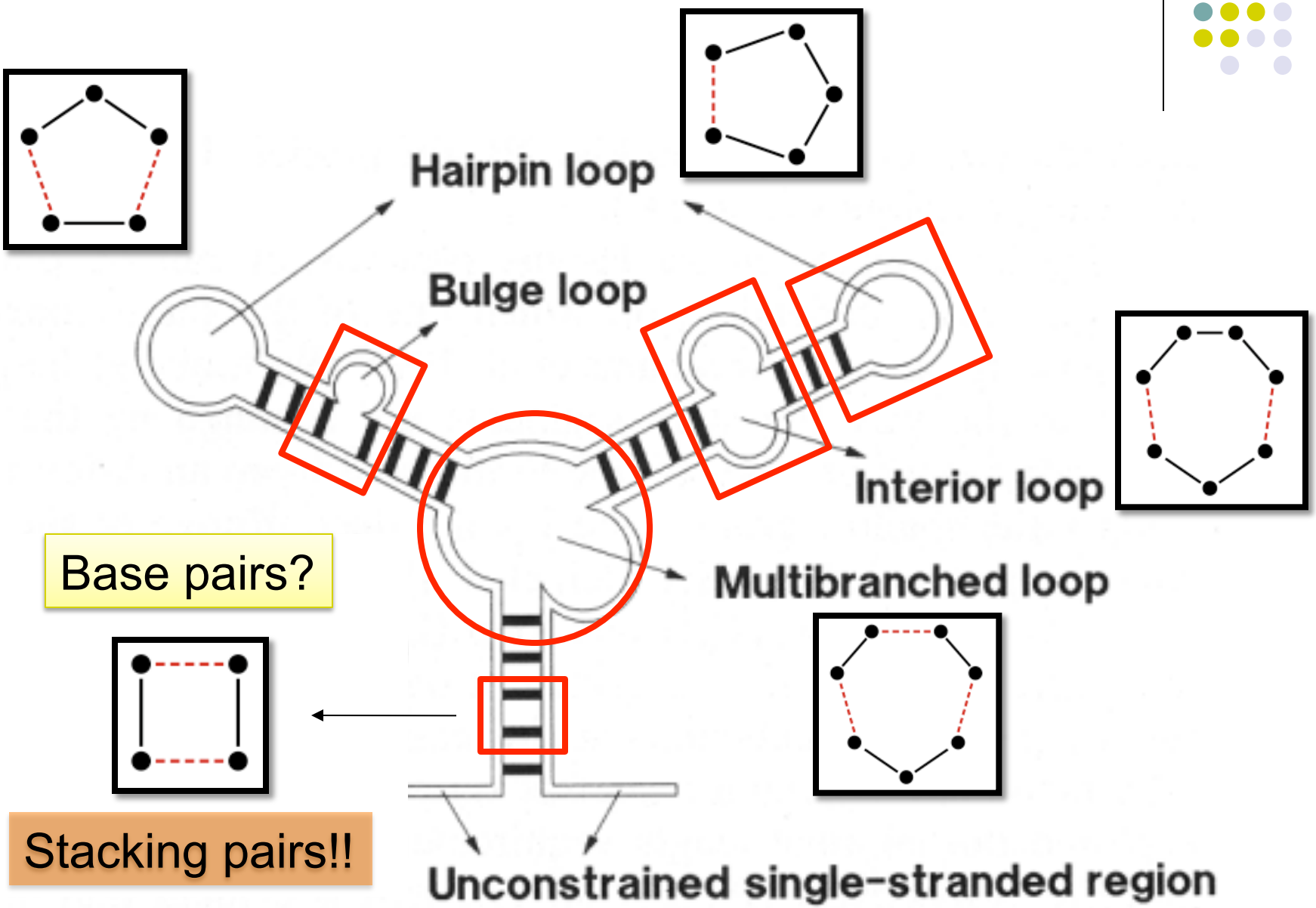
But we need:

- *to define what are the important structural features that has to be evaluated.*
- *to keep the energy contribution local in order to allow a divide-and-conquer aproach (fast).*

RNA secondary structure elements

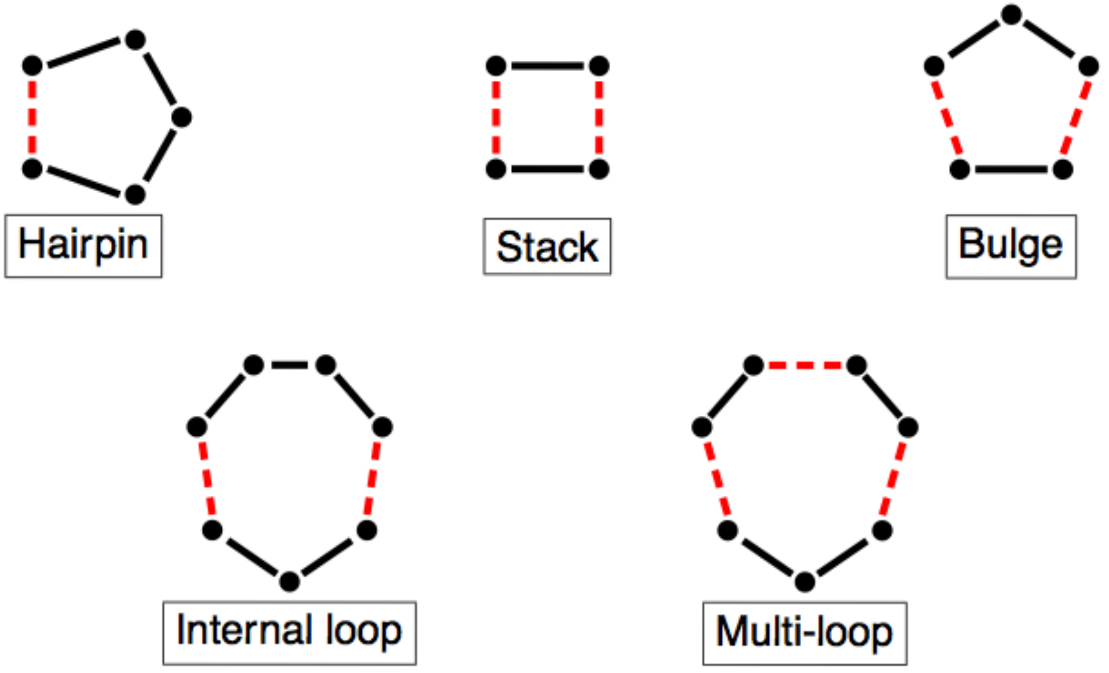


Loop decomposition



RNA secondary structure description

A secondary structure can be decomposed in a sequence of loops:

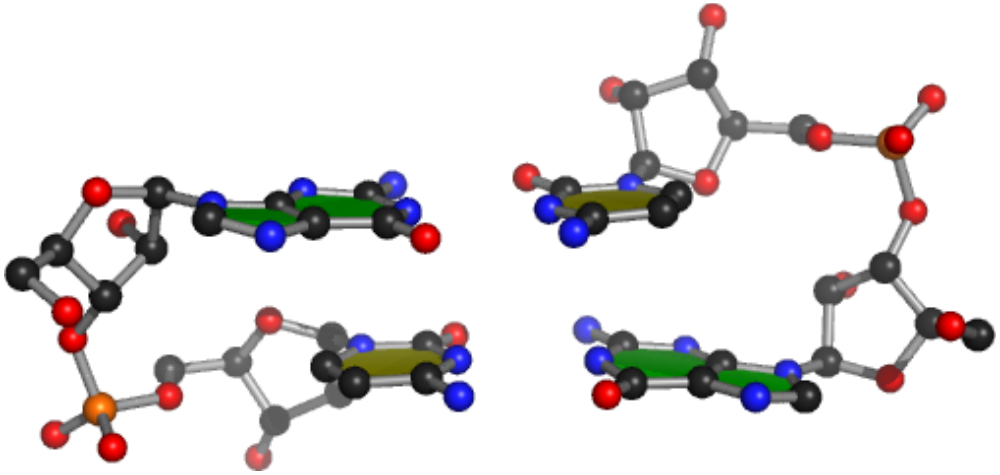


: Sequence neighbors

: Spatial neighbors

Stacking base pairs

Base stacking interactions between the pi orbitals of the bases' aromatic rings contribute to stability. GC stacking interactions with adjacent bases tend to be more favorable.



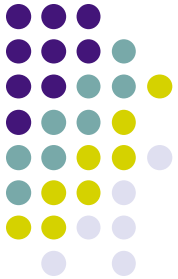
Note: Stacking energy are orientated.

5’ - CG - 3’
3’ - GC - 5’

≠

5’ - GC - 3’
3’ - CG - 5’

RNA nearest neighbor energy model



Unpaired state ↔ Structure i

$$K_i = \frac{[\text{Structure i}]}{[\text{Unpaired state}]} = e^{-\Delta Gi/RT}$$

Structure i ↔ Structure j

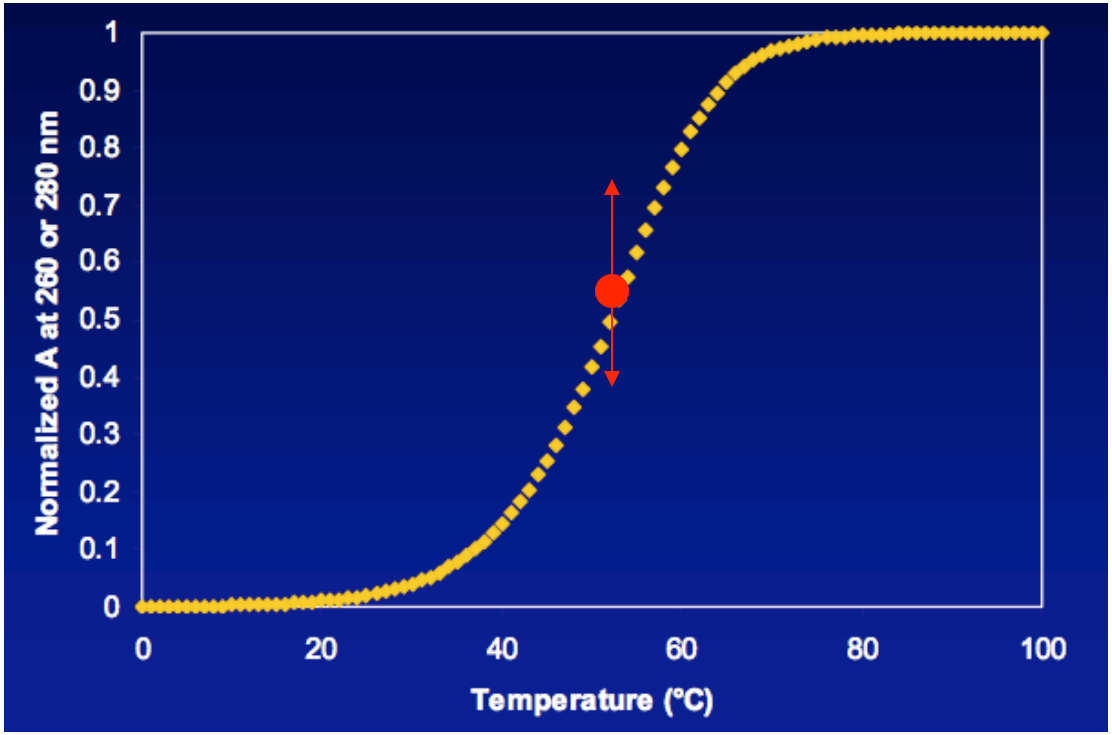
$$\frac{[\text{Structure i}]}{[\text{Structure j}]} = K_i/K_j = e^{-(\Delta Gi-\Delta Gj)/RT}$$

The Gibbs free energy ΔG quantify the favorability of a structure at a given temperature.

ΔG is experimentally estimated from optical melting curves.

Optical melting curves

The UV-absorbance melting curves estimate the number of base pair in the duplex. At the melting point the change in Gibbs free energy (ΔG) is zero. 50% of the oligonucleotide and its perfect complement are in duplex.The melting temperature correspond to the inflexion point of the curve fitted to the 2 state model (Xia et al., 1999).



Here: T_m = Melting temperature = 52°C

RNA nearest neighbor energy model

Hairpin : positive destabilizing energy of a hairpin with k unpaired bases. Bonuses for tri- and tetra- and GGG-loops.

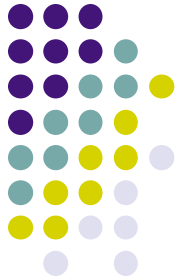
Stack : negative stabilizing energy of an additional stacked base pair.

Bulge : positive energy of a bulge with k unpaired bases. Add stacking energy if size is 1.

Internal loop : positive energy of an interior loop with k and m unpaired bases in bulges. Special cases of 1x1, 1x2, 2x1 and 2x2 internal loops. Penalty for the asymmetry.

Multi-loop : *linear energy approximation* $\alpha + \beta \cdot N_u + \gamma \cdot N_h$, where N_u is the number of unpaired bases and N_h the index of the multi-loop (i.e. the number of connected helices).

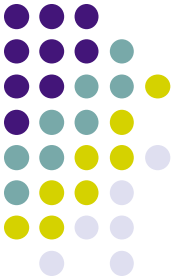
RNA nearest neighbor energy model



Other Parameters:

- Dangles (unpaired nucleotides at stem extremities).
- Extrapolation for large loops based onpolymer theory.
- Internal, bulge or hairpin-loops > 30: $dS(T)=dS(30)+\langle param \rangle \ln(n/30)$.
- Terminal AU penalty.
- GAIL rule (asymmetric interior loop rule).
- Coaxial stacking.
- Logarithmic energy function for multi-loop (break the dynamic programming scheme)

Zuker Algorithm



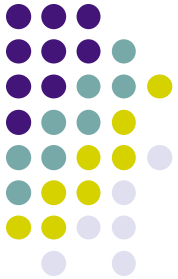
Goal: Computing the minimum free energy secondary structure.

Can be achived using dynamic programming (Zuker-Stiegler,81)

Dynamic table:

- E_h : first and the last paired nucleotides base-pair together. Example : $...(***)..$, where . denotes an unpaired position, *** denotes any valid substructure.
- E_h^* : leftmost and rightmost nucleotides of the sub-sequence base pair together. Example : $(***)$
- E_e : At least 2 stems occur in an exterior loop. Example : $...(***)..(***)...$
- E_1^m : Same as E_h, except that a penalty for unpaired bases occurring in a multi-loop is added for each nucleotide occurring outside the stem.
- E_m^2 : At least 2 stems appear in a multi-loop. In this case,a penalty is added for unpaired bases outside each stem.

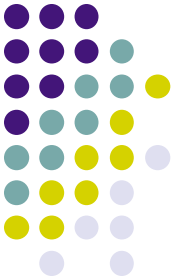
Zuker Algorithm



Energy functions:

- $Ehairpin(i, j)$: Energy of a Hairpin closed at index (i, j). Includes all bonuses.
- $Eloop(i, m, n, j)$: energy of a loop of index 2. Includes :
  - Stacks ($m = i + 1, n = j - 1$),
  - Bulges ($m - i > 1$ xor $j - n > 1$),
  - Internal loops ($m - i > 1$ and $j - n > 1$).
- Multi-loop energy parameters :
  - α : affine constant,
  - β : unpaired nucleotide penalty,
  - γ : helix penalty,
- $Edangle(i, j)$: energy of a dangle.

Zuker Algorithm



Algorithm 3

for $d = \theta$ *to* $n - 1$

  for $i = 0$ *to* $n - d$

    $j = i + d$;

    for $r = i$ *to* $j - theta - 1$

      if basepair(i,j) :

        if $r = i$:

$$E_h^*(i,j) = Ehairpin(i,j)\,;$$
$$E_h^*(i,j) = \min(E_h^*(i,j), E_m^2(i+1,j-1) + \alpha + \gamma + Edangle(j,i))$$

for (m,n) s.t. $i < m < n < j$ *and basepair(m,n) :*

$$E_h^*(i,j) = \min(E_h^*(i,j), Eloop(i,m,n,j) + E_h^*(m,n))$$

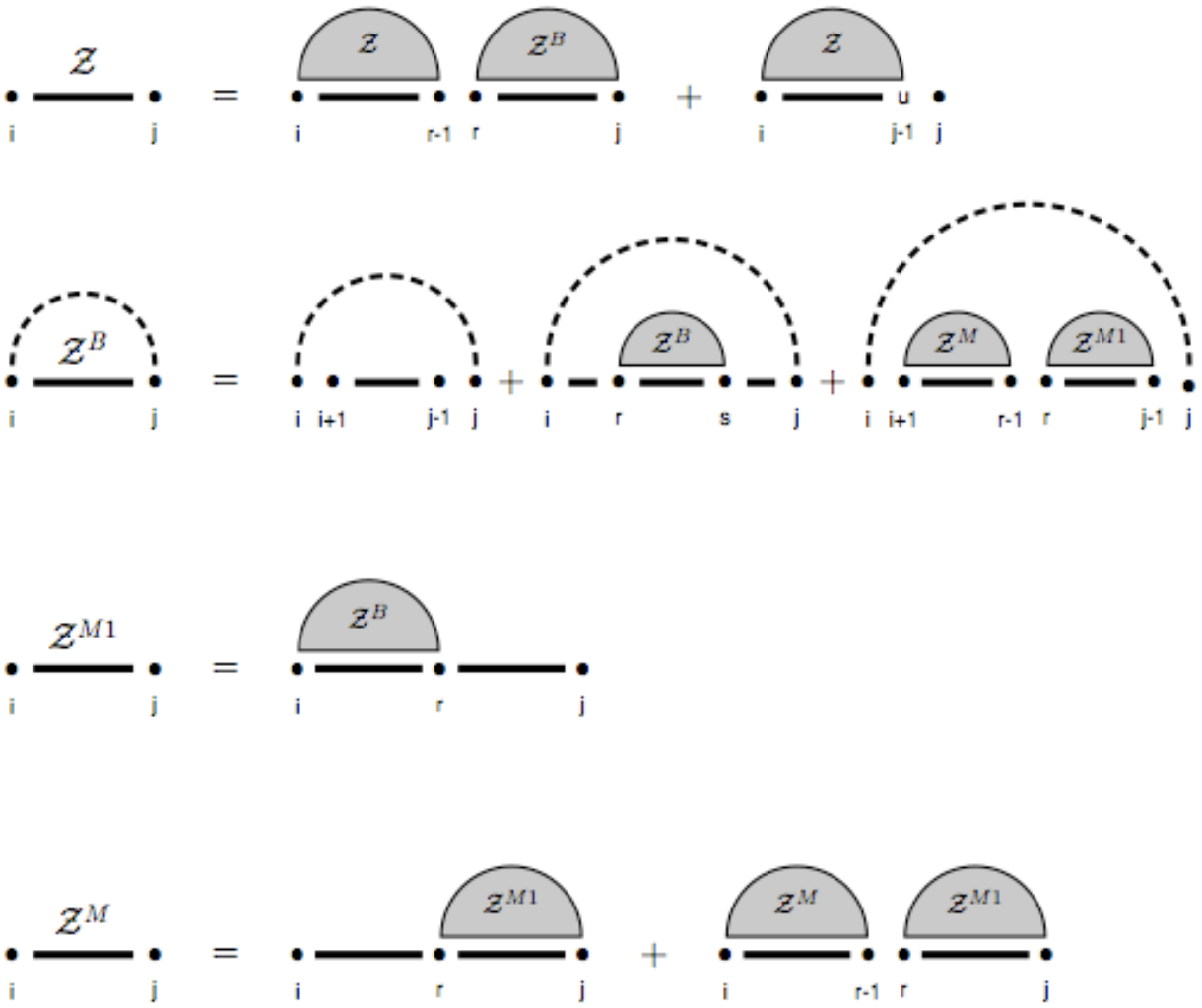
$$E_m^1(i,j) = E_h(i,j) = E_h^*(i,j) + Edangle(i,j)$$

        else :

$$E_h(i,j) = \min(E_h(i,j), E_h^*(r,j) + Edangle(i,j))$$
$$E_m^1(i,j) = \min(E_m^1(i,j), E_h^*(r,j) + Edangle(r,j) + (r-i)\cdot\beta)$$
$$E_m^2(i,j) = \min(E_m^2(i,j), E_m^1(i,r-1) + E_h^*(r,j) + Edangle(r,j) + 2\cdot\gamma)$$
$$E_m^2(i,j) = \min(E_m^2(i,j), E_m^2(i,r-1) + E_h^*(r,j) + Edangle(r,j) + \gamma)$$
$$E_e(i,j) = \min(E_e(i,j), E_h(i,r-1) + E_h^*(r,j) + Edangle(r,j))$$
$$E_e(i,j) = \min(E_e(i,j), E_e(i,r-1) + E_h^*(r,j) + Edangle(r,j))$$

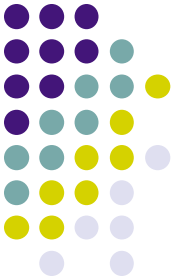
$$E_e(i,j) = \min(E_e^1(i,j), E_e(i,j-1))$$
$$E_m^2(i,j) = \min(E_m^2(i,j), E_m^2(i,j-1) + \beta)$$
$$E_h(i,j) = \min(E_h(i,j), E_h(i,j-1))$$
$$E_m^1(i,j) = \min(E_m^1(i,j), E_m^1(i,j-1) + \beta)$$

Zuker Algorithm: Feyman Diagrams



Schematic representation of the recursive equations.

Zuker Algorithm



- The RNA minimum free energy (m.f.e.) is $\min(E_h(1,N),E_e(1,N))$.
- The m.f.e. structure can be obtained by backtracking.

Warning: this (simplified) algorithm does not check when dangle penalty must be applied or not.

This algorithm is implemented in *UNAfold* (previously Mfold), the *Vienna RNA package* (RNAfold) and *RNAstructure* (for windows).