Basics of RNA structure prediction

- Two primary methods of structure prediction
  - Covariation analysis/Comparative sequence analysis
    - Takes into account conserved patterns of basepairs during evolution (2 or more sequences).
    - Pairs will vary at same time during evolution yet maintaining structural integrity
    - Manifestation of secondary structure
  - Minimum Free-Energy Method
    - Using one sequence can determine structure of complementary regions that are energetically stable

Comparative Sequence Analysis

- **Molecules with similar functions and different nucleotide sequences will form similar structures.**

- **Predicts secondary and tertiary structure from underlying sequence.**

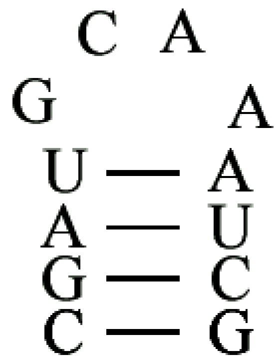
- **Correctly identifies high percentage secondary structure pairings and a smaller number of tertiary interactions.**

- **Primarily a manual method**

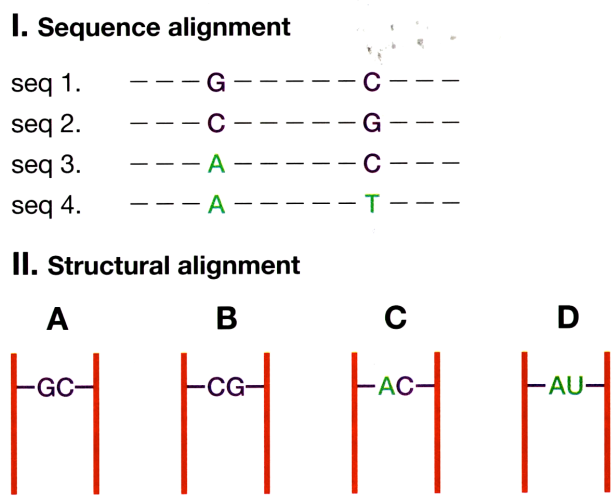
Positional Covariation

- Helix is formed from two sets of sequences that are not identical.

C G A U (G C A A) A U C G



- Search for positions that co-vary.



- Positions that co-vary with one another are possible pairing partners.

Support for Comparative Models?

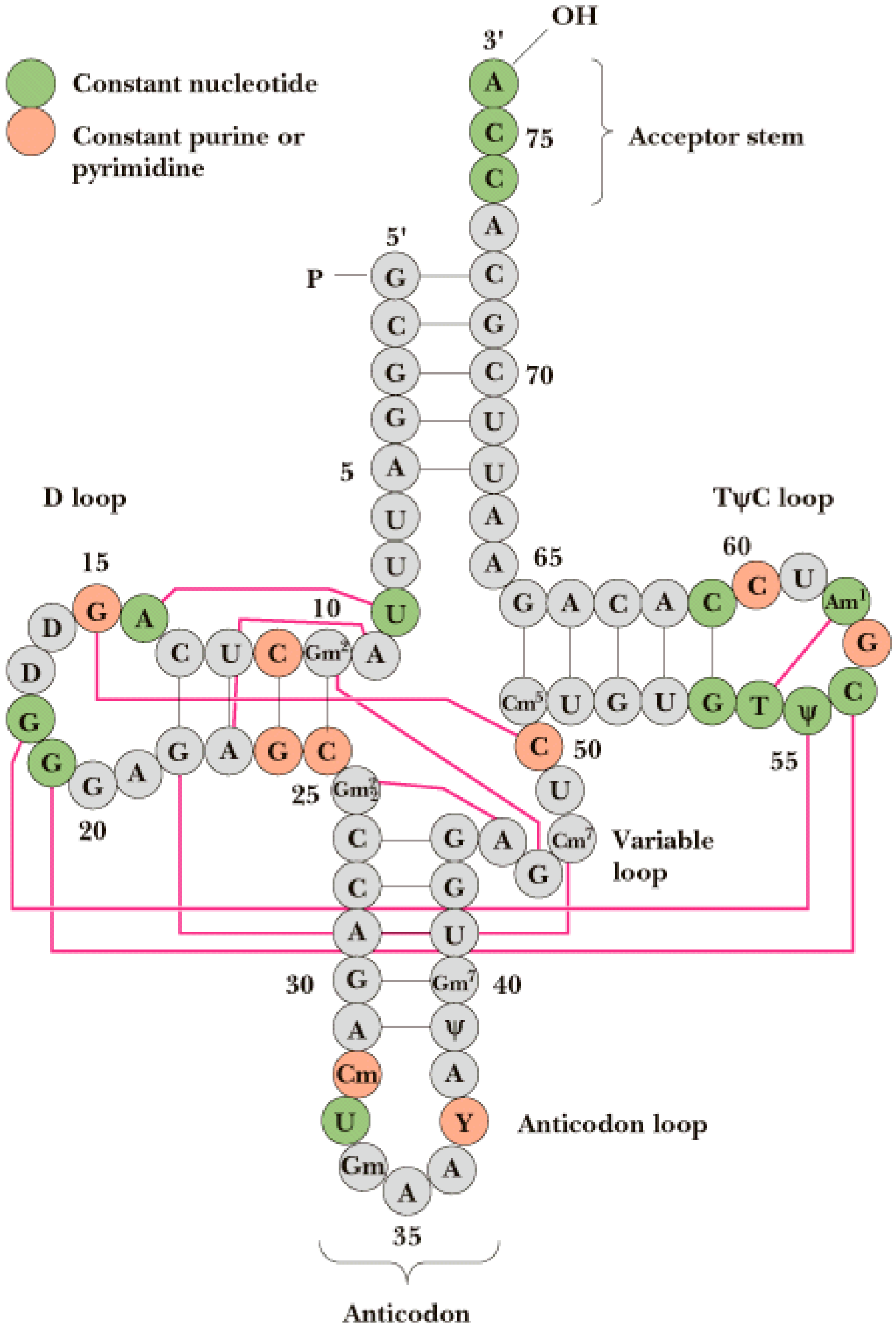
- **Comparative vs. Experimental**
  - **Estimate that ~98% of pairings in current comparative model will be in the crystal structure**

- **Interactions not easily identified**
  - **Tertiary base-pairings**
  - **Aim to predict all interactions with comparative analysis**

Thus, comparative sequence analysis predicts almost all of the secondary structure base-pairs and some tertiary pairings present in crystal structures.

tRNA structure

Tertiary pair or contact



Comparative sequence analysis

The 2D of all structured RNAs have been obtained by this method :
tRNAs, rRNAs, RNaseP, group I and group II introns, snRNAs, SRP RNAs, etc.

SANKOFF's problem : align and derive the 2D structure from a set of non-aligned sequences : NP-complete !

Working hypothesis

The native secondary structure is the one with the minimum free energy.

Basic Model

- RNA linear structure: $R=r_1\, r_2\, .\, .\, .\, r_n$ from {A,C,G,U}
- RNA secondary structure: pairs (r_i,r_j) such that $0<i<j<n+1$.
- Goal: secondary structures with minimum free energy.

Implementing Model Restrictions

- No knots: pairs (r_i, r_j) and (r_k, r_l) such that $i<k<j<l$. RNA does contain knots.
- No "close" base pairs: $j-i>t$ for some $t>0$.
- Complementary base pairs: A-U, C-G
  with the wobble pair GoU

Tinoco-Uhlenbeck postulate

- Assumption: The energy of each base pair is independent of all of the other pairs and the loop structure.
- Consequence: Total free energy is the sum of all of the base pair free energies.

Independent Base Pairs
Basic Approach

- Use solutions for smaller strings to determine solutions for larger strings.
- This is **precisely** the kind of decoupling required for dynamic programming algorithms to work.

Independent Base Pairs Notation

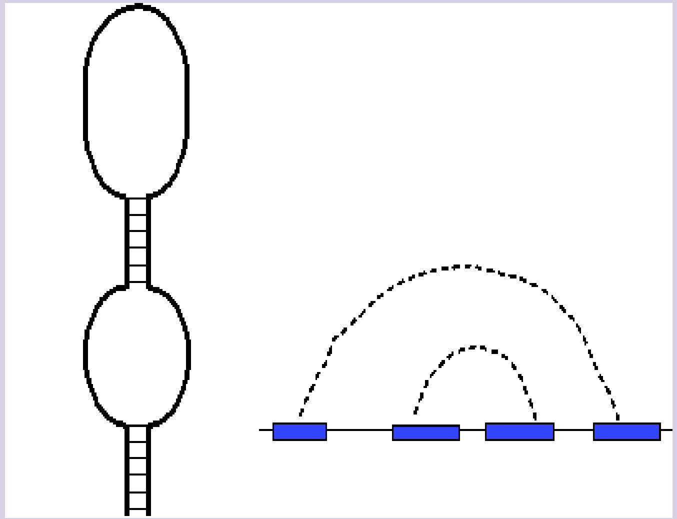
- $a(r_i,r_j)$ – the free energy of a base pair joining r_i and r_j.
- $S_{i,j}$ – The secondary structure of the RNA strand from base r_i to base r_j. Ie, the set of base pairs between r_i and r_j inclusive.
- $E(S_{i,j})$ – The free energy associated with the secondary structure $S_{i,j}$.
- We define $a(r_i,r_j)$ large when constraints are violated.

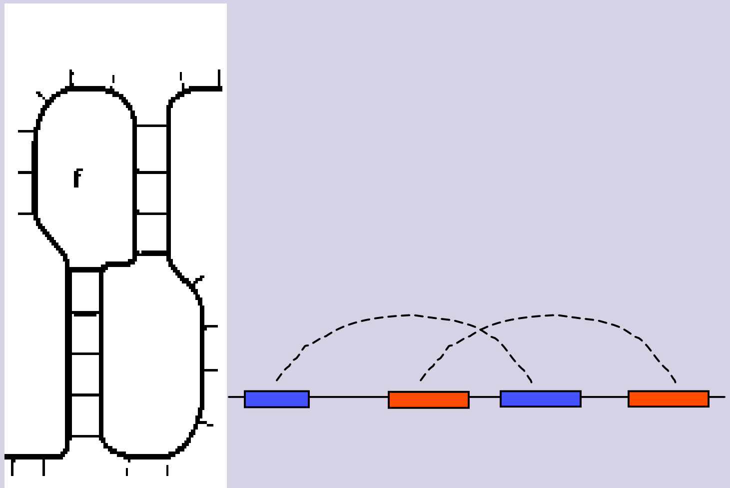
Independent Base Pairs: Calculating Free Energy

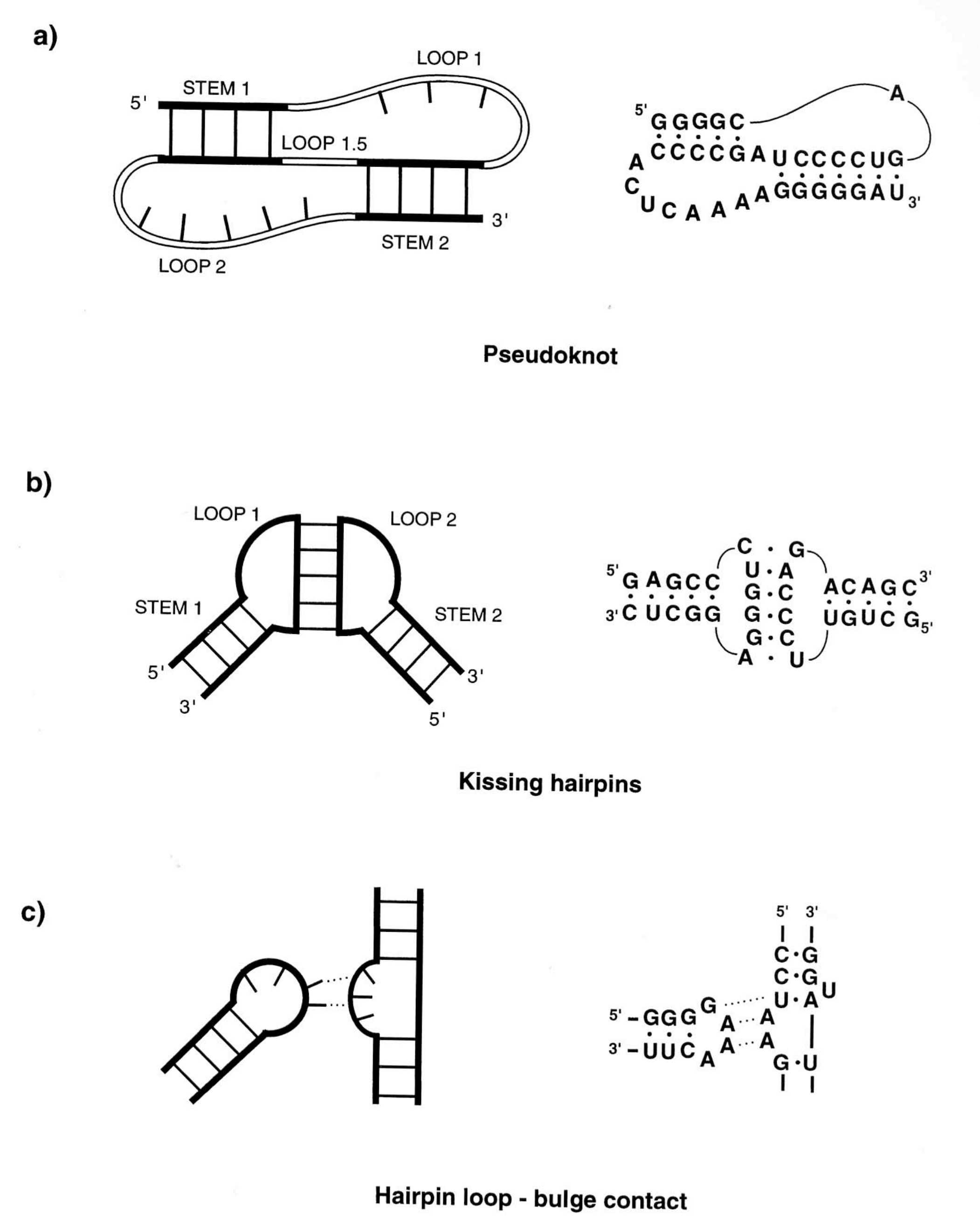
- Consider the RNA strand from position i to j.
- Consider whether r_j is paired
- If r_j is paired, $E(S_{i,j})=E(S_{i,k-1})+a(k,j)+E(S_{k+1,j-1})$ for some $i-1<k<j$
- If r_j isn't paired, then $E(S_{i,j})=E(S_{i,j-1})$

Non-canonical pairs and pseudoknots

- In addition to A-U and G-C pairs, **non-canonical pairs** also occur. Most common one is G-U pair, the wobble pair.
- G-U is thermodynamically favourable as Watson-Crick pairs (A-U, G-C) .
- Base pairs almost always occur in nested fashion. Exception: **pseudoknots**.







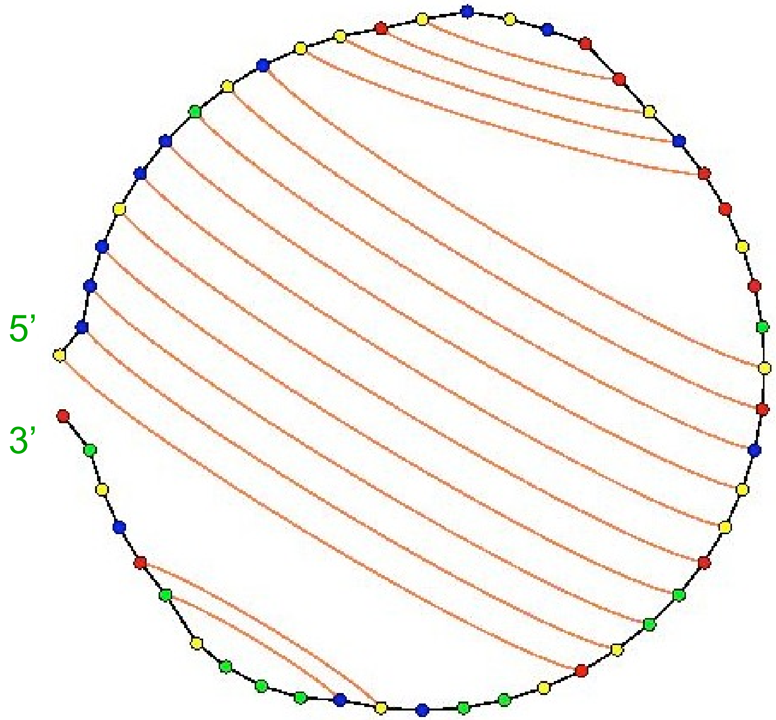
RNA Tertiary Structure

- Do not obey "parentheses rule"

Computational Complexity

Without Pseudoknot

GUUUGUUAGUGGCGUGUCCGUCCGCA
GCUGGCAAGCGAAUGUAAAGACUGAC



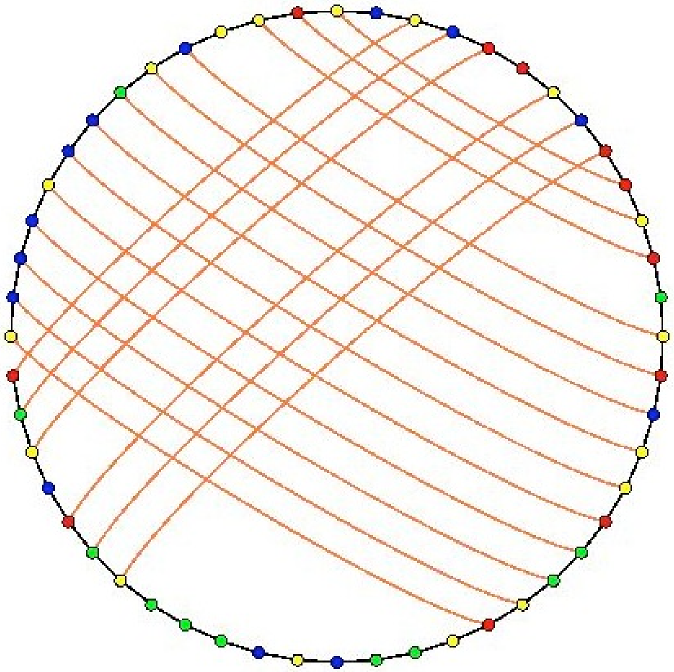
Rainbow constraint:

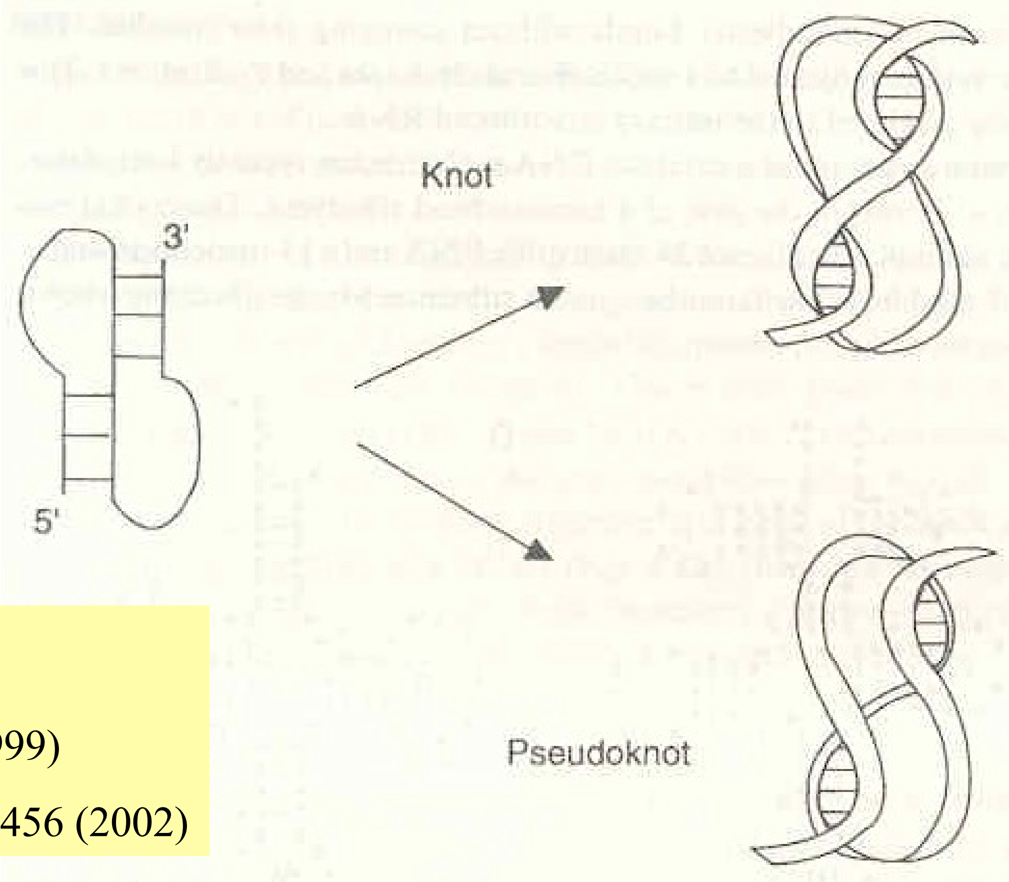
any two pairs $i<j$ and $i'<,j'$

satisfy $i<i'<j'<j$ or $i'<i<j<j'$

computational steps: N^3

H-Pseudoknot



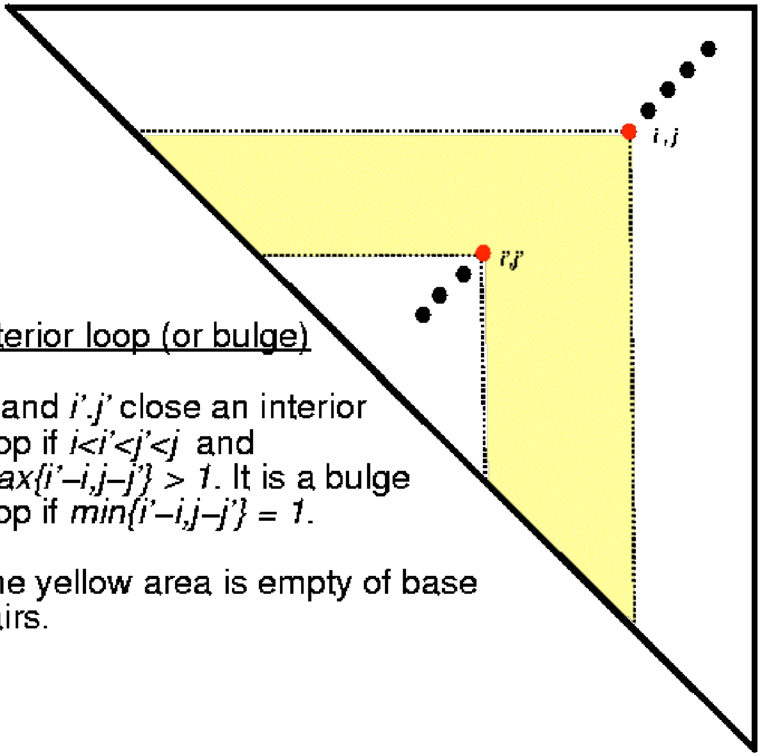


Exact: at least N^6

Rivas and Eddy, JMB **285**, 2053 (1999)

Orland and Zee, Nucl. Phys. B **620**, 456 (2002)

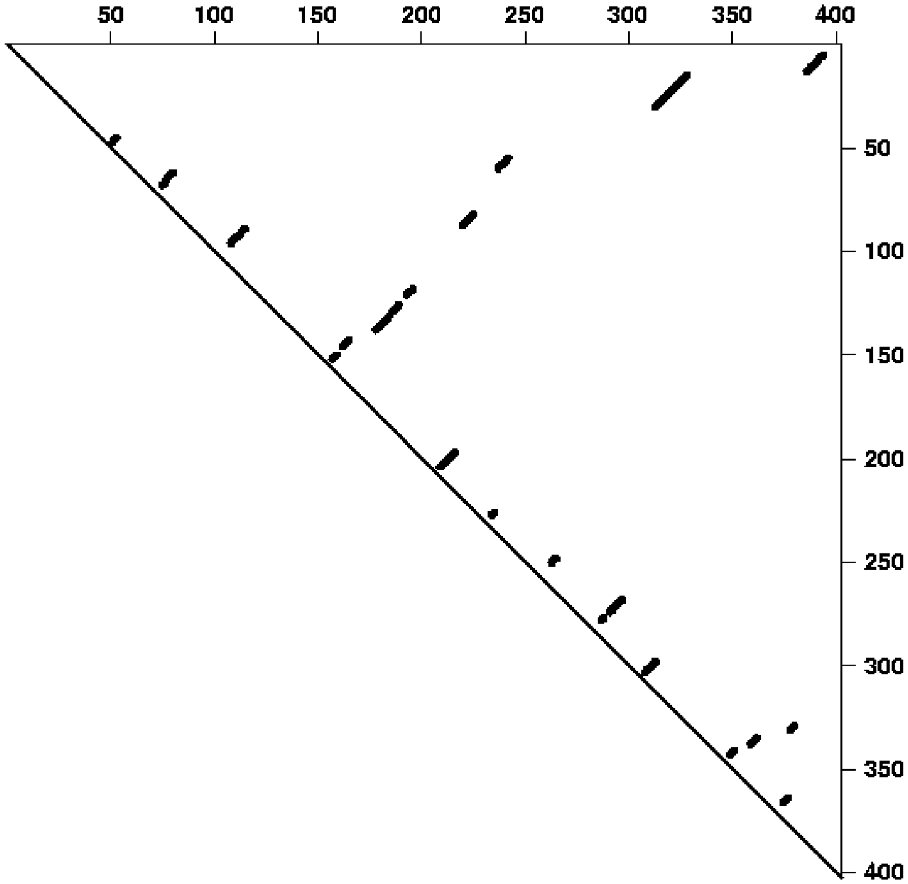
Dot plot

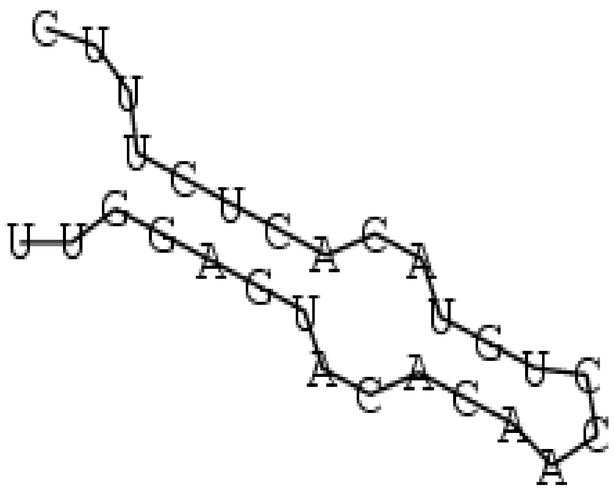


<u>Interior loop (or bulge)</u>

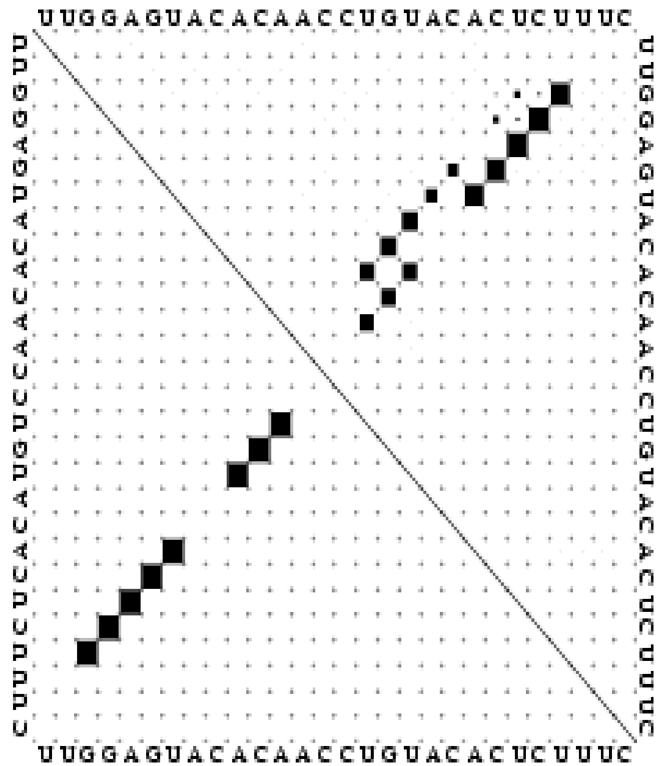
$i.j$ and $i'.j'$ close an interior loop if $i<i'<j'<j$ and $max\{i'-i,j-j'\} > 1$. It is a bulge loop if $min\{i'-i,j-j'\} = 1$.

The yellow area is empty of base pairs.




C U U U C U C A C A U C U C C A A C A C A U G A C G U U

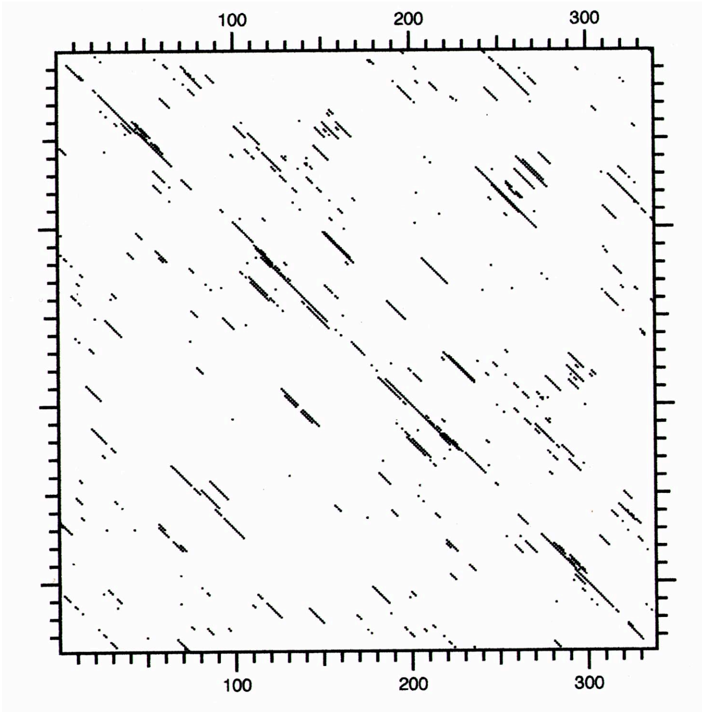
rna.ps


UUGGAGUACACAACCUGUACACUCUUUC

dot.ps

Minimum Free-Energy Method

- Searching for structures with stable energies
- First a dot matrix analysis is carried out to highlight complementary regions (diagonal indicates succession of complementary nucleotides)
- The energy is then calculated for each predicted structure by summing negative base stacking energies



Free energy values for RNA structure

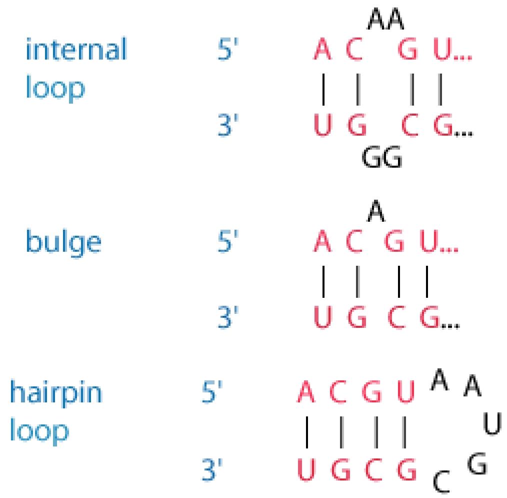
- Complementary regions are evaluated using the dynamic programming algorithm to predict the most energetically stable molecule

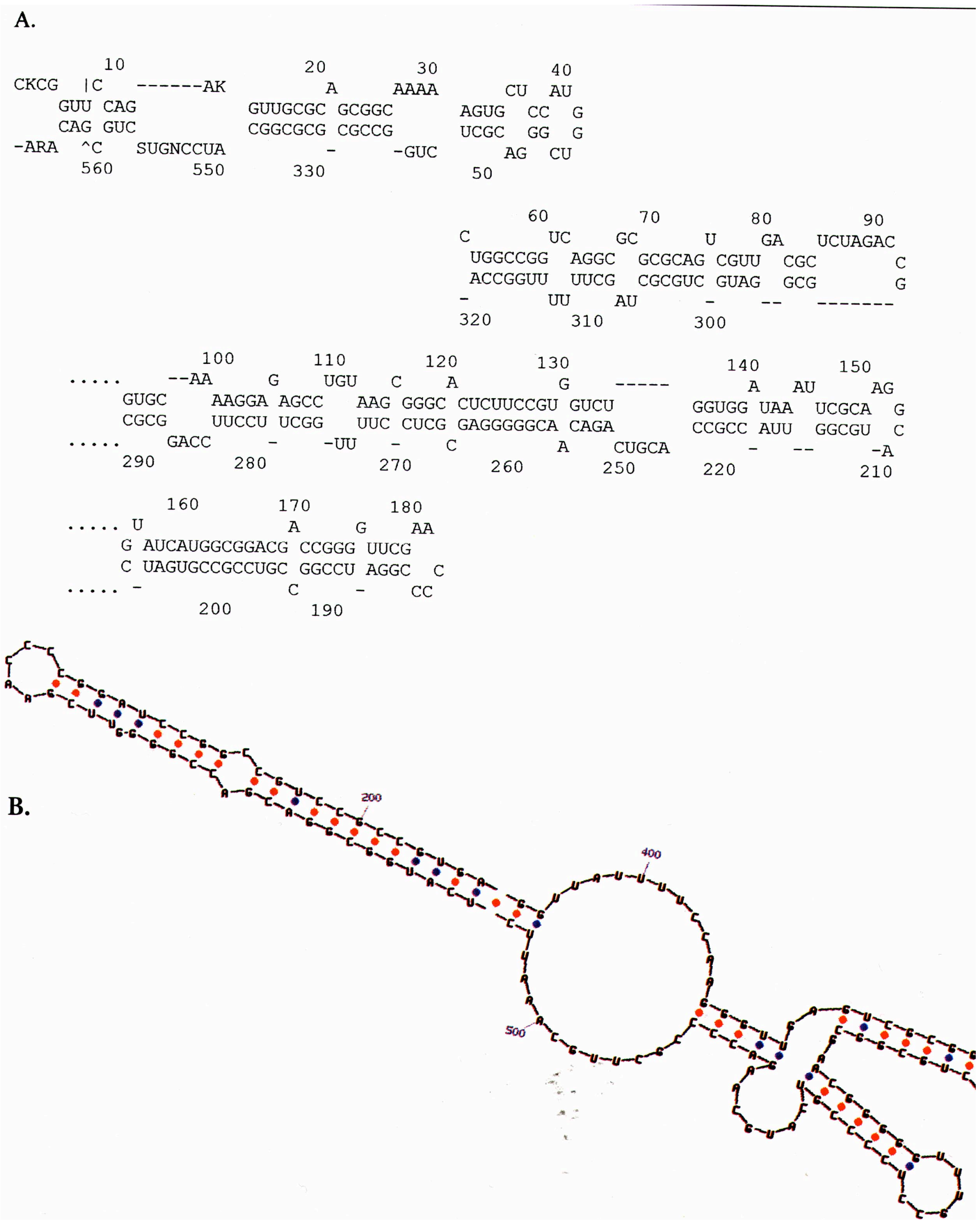
A. Stacking energies for base pairs

	A/U	C/G	G/C	U/A	G/U	U/G
A/U	−0.9	−1.8	−2.3	−1.1	−1.1	−0.8
C/G	−1.7	−2.9	−3.4	−2.3	−2.1	−1.4
G/C	−2.1	−2.0	−2.9	−1.8	−1.9	−1.2
U/A	−0.9	−1.7	−2.1	−0.9	−1.0	−0.5
G/U	−0.5	−1.2	−1.4	−0.8	−0.4	−0.2
U/G	−1.0	−1.9	−2.1	−1.1	−1.5	−0.4

B. Destabilizing energies for loops

Number of bases	1	5	10	20	30
internal	–	5.3	6.6	7.0	7.4
bulge	3.9	4.8	5.5	6.3	6.7
hairpin	–	4.4	5.3	6.1	6.5




A.
10 20 30 40
CKCG |C ------AK A AAAA CU AU
GUU CAG GUUGCGC GCGGC AGUG CC G
CAG GUC CGGCGCG CGCCG UCGC GG G
-ARA ^C SUGNCCUA - -GUC AG CU
560 550 330 50
60 70 80 90
C UC GC U GA UCUAGAC
UGGCCGG AGGC GCGCAG CGUU CGC C
ACCGGUU UUCG CGCGUC GUAG GCG G
- UU AU - -- -------
320 310 300
100 110 120 130 140 150
..... --AA G UGU C A G ----- A AU AG
GUGC AAGGA AGCC AAG GGGC CUCUUCCGU GUCU GGUGG UAA UCGCA G
CGCG UUCCU UCGG UUC CUCG GAGGGGGCA CAGA CCGCC AUU GGCGU C
..... GACC - -UU - C A CUGCA - -- -A
290 280 270 260 250 220 210
160 170 180
..... U A G AA
G AUCAUGGCGGACG CCGGG UUCG
C UAGUGCCGCCUGC GGCCU AGGC C
..... - C - CC
200 190
B.
200
400
500

Example

Partition function

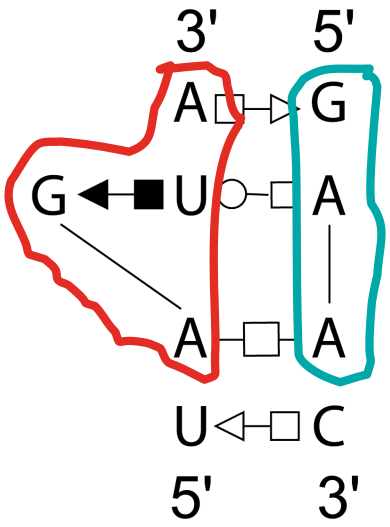
- Definition:

$$Q = \sum_{S \in \mathcal{S}} e^{-\frac{\Delta G_S}{RT}}.$$

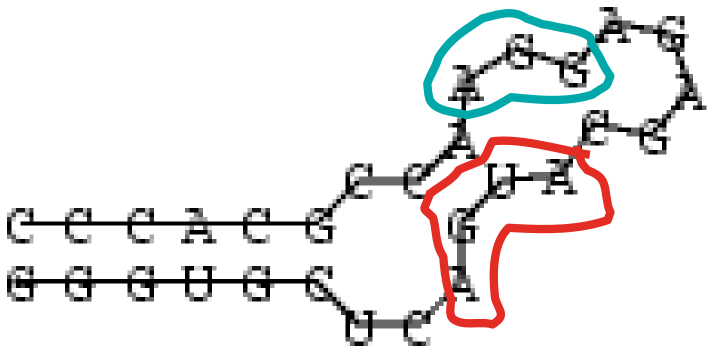
- This is a weighted counting of all structures.
- The lower the free energy, the higher the weighting.
- According to statistical mechanical theory, this Boltzmann weighting gives the probability density for every folding.

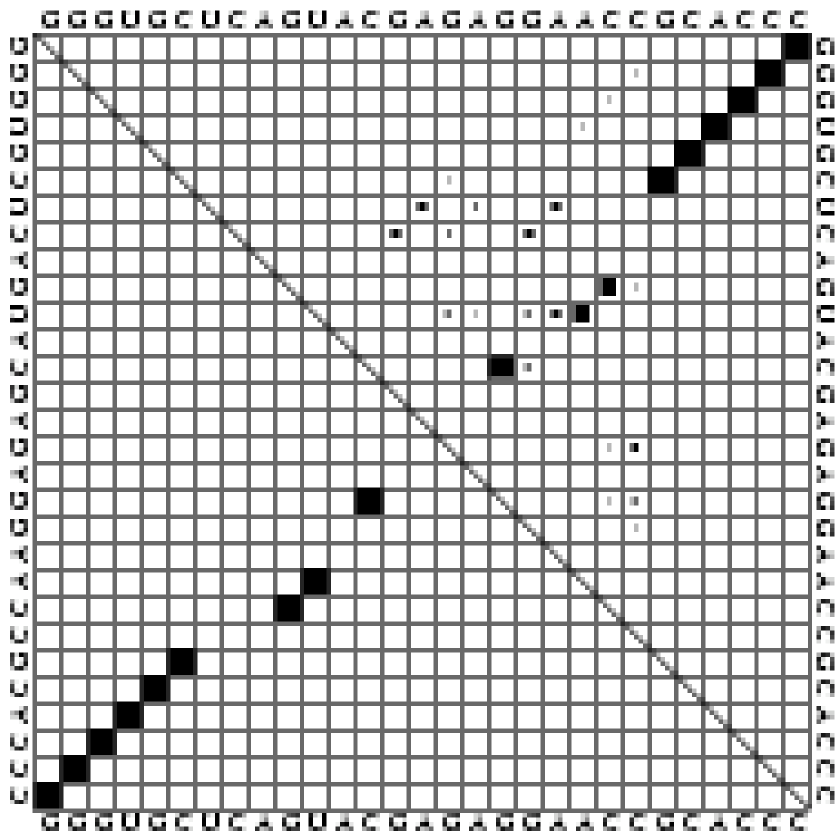
$$\Pr(S) = \frac{e^{-\frac{E(S)}{RT}}}{Q_{1,n}}$$

- Partition function does not predict a secondary structure but can calculate the probability for a certain base pair to form.



Loop E motif is a continuous Stack of non-Watson-Crick pairs





Some webpages to check out

- Comparative RNA Web site (CRW)
  - http://www.rna.icmb.utexas.edu
- MFOLD minimum energy RNA
  - http://bioinfo.math.rpi.edu/~zukerm/rna/
- RNA world
  - http://www.imb-jena.de/RNA.html
- RNA structure database
  - http://www.rnabase.org/
- Database of ribosomal subunit sequences
  - http://rrna.uia.ac.be/

Inverse folding

Another direction in sequence design is designing a sequence that folds into a given secondary structure. This problem is called *inverse folding*, because it is the inverse of the problem of finding the secondary structure of a sequence with the minimum free energy. The inverse folding problem is to find a sequence whose minimum energy structure coincides with the given one

Inverse folding