RNA

- two topics
  - structure prediction
  - why it may not matter
- why is RNA so fashionable ?
  - enzymatic activity (RNAzymes, hammerhead, ribosome)
  - specific ligand binding
    - regulators, riboswitches
  - temperature sensors
  - ubiquitous transcription
  - nobel prize for ribozome

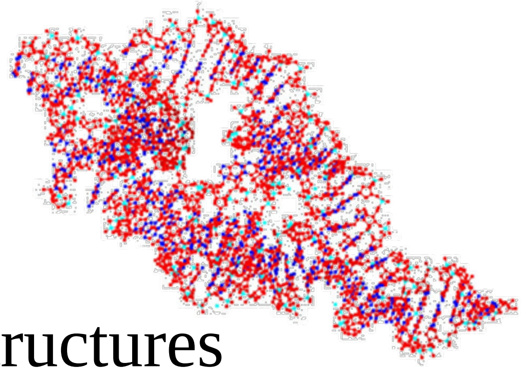
  - first life on earth ?

comparison to proteins

Analogy to proteins

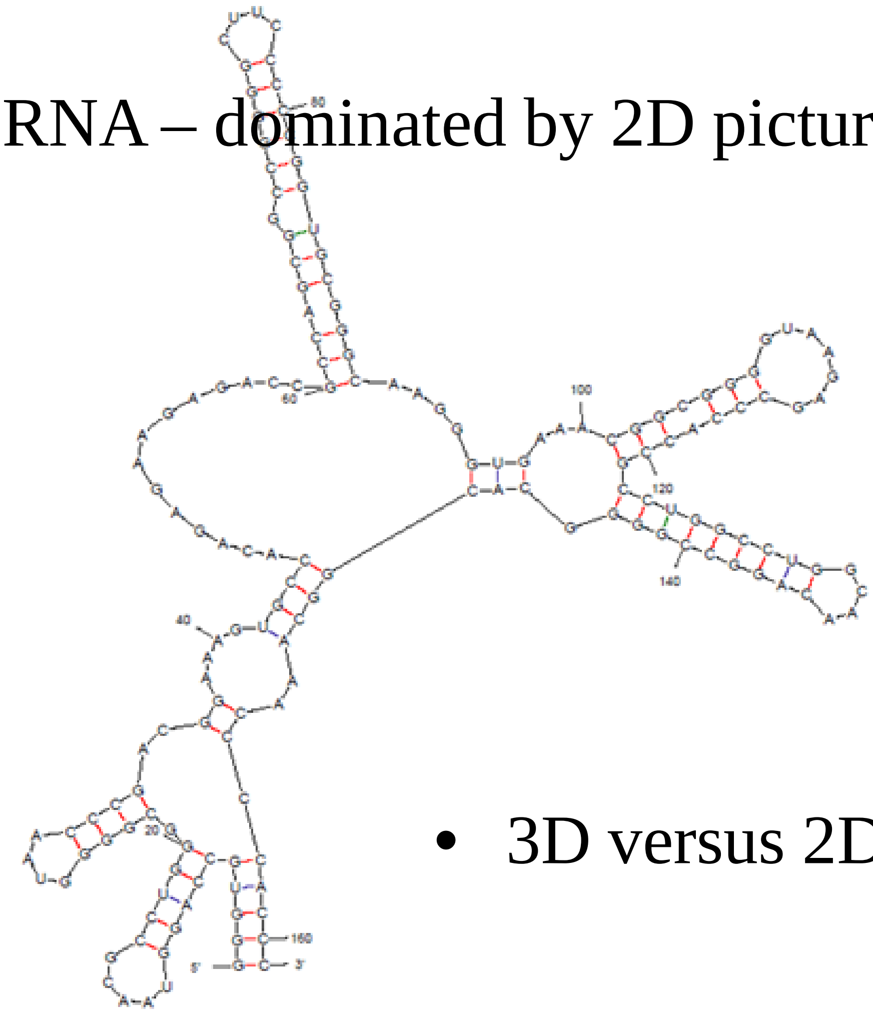
- Proteins
    - common belief – unique structure for sequence
    - 20 amino acids, many specific interactions
        - hydrophobic, charged, big, small, …
        - hydrophobic core
    - 6.8 ×10^4 structures in databank
- RNA
    - < 10^3 structures in databank
    - 4 bases
        - 2 bigger, 2 small (A, G, C, U)
    - less specificity ? fewer unique structures

2D and 3D

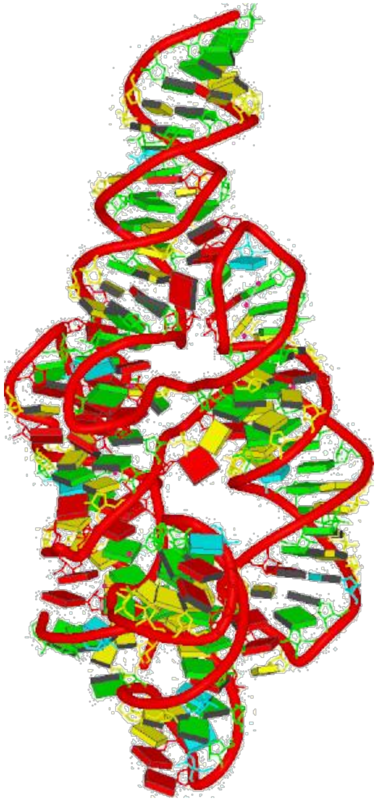


- proteins – usually talk of sequences or 3D structures

- RNA – dominated by 2D pictures



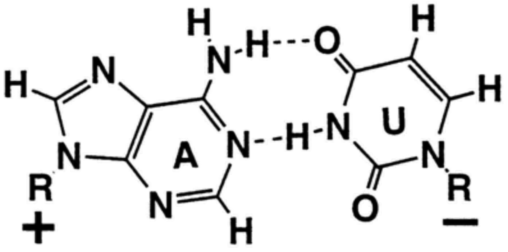
- 3D versus 2D (1u9s)

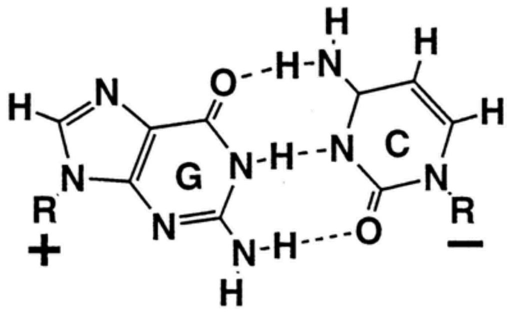


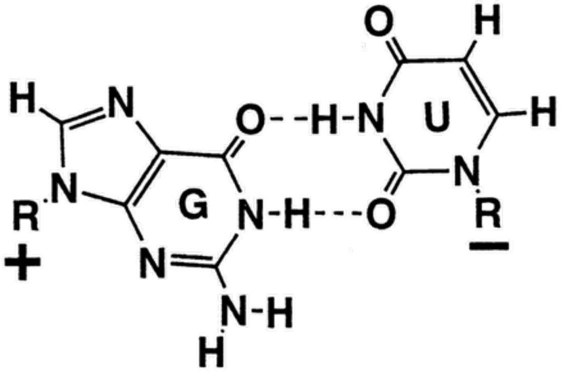
2D why of interest ?

1. computationally tractable
2. historic – belief that nucleotides are
   - dominated by classic (Watson-Crick) H-bonds

- later – GU wobble pairs

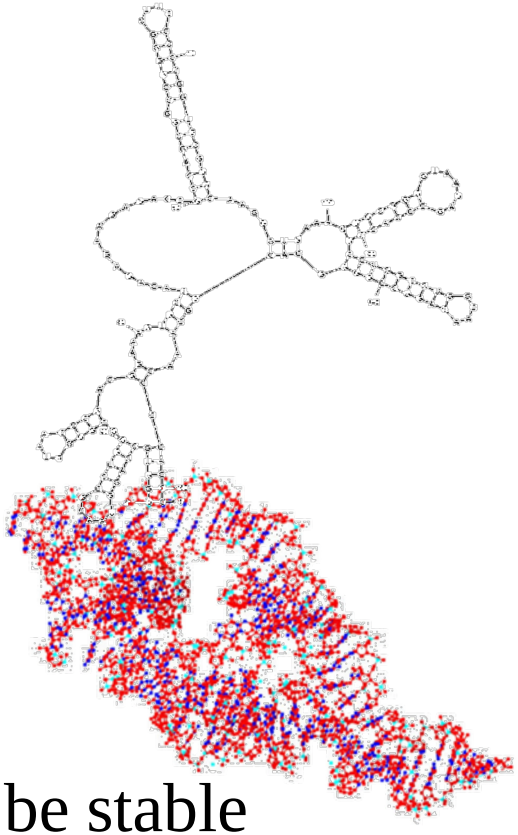






from Burkard, M.E., Turner, D.H., Tinoco Jr., I., in The RNA World, 2nd Edn, eds Gesteland, RF, Cech, TR., Atkins, JF Cold Spring Harbor Laboratory Press (1999)

2D why of interest ?

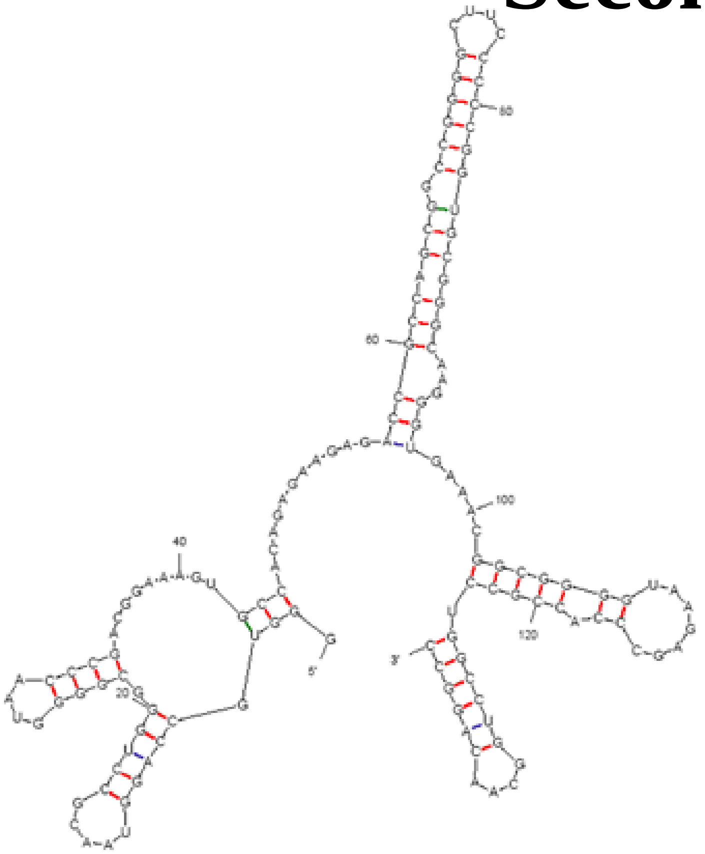


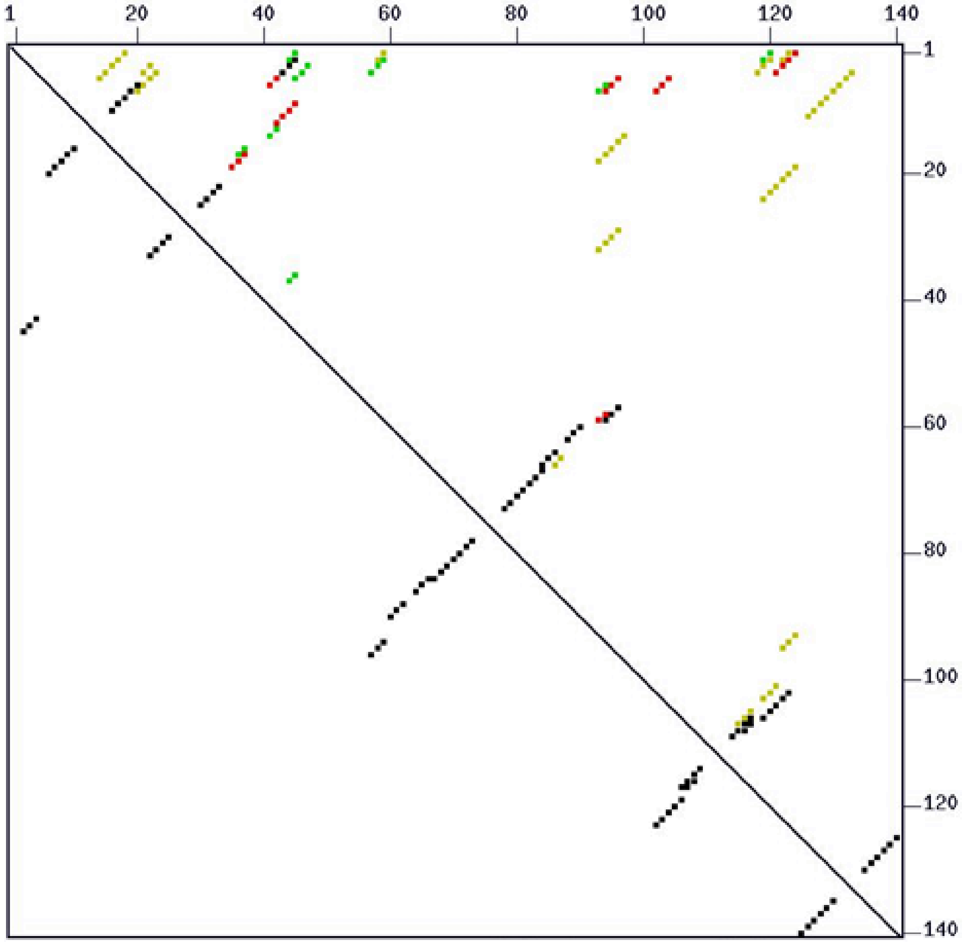
3. Claim - RNA folds hierarchically

   nearby bases fold first, later overall structure

- evidence not clear
- much contrary evidence in protein world
- plausible in RNA world ?
  - RNA double strand helices are believed to be stable
  - contrast with proteins – isolated α-helices and β-strands are not stable in solution
- useful ?
  - if true, then 2D (H-bond pattern) prediction is really the first step to full structure prediction

Secondary Structure





- same features in both plots
    - look for long helix 57-97, bulges in long helix

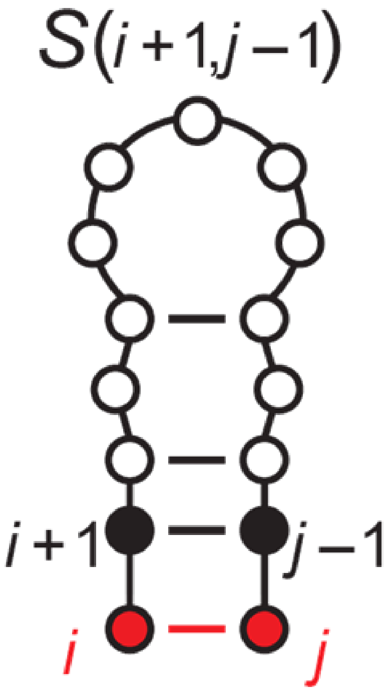
Predicting secondary structure

- Ingredients
  - scoring scheme
    - more base pairs – better
    - more sophisticated later
  - some restrictions on ordering of pairs (more later)

- dynamic programming method
  - 1 step more than sequence alignments

hairpins

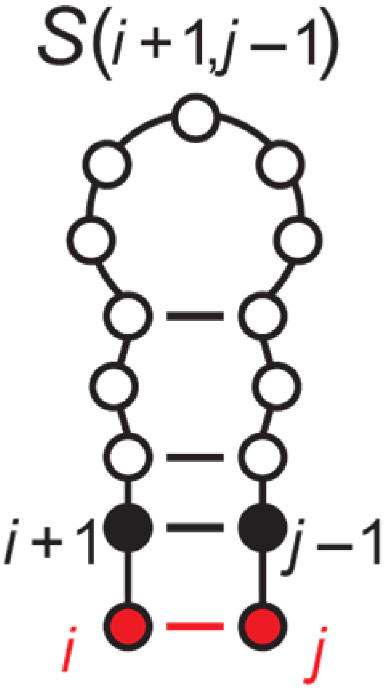
- start by looking for best possible hairpin
- idea
  - if we know the structure of the inner loop
    - we can work out the next
  - if we know the black parts
    - we can decide what to do with the red *i* and *j*



picture from Eddy, S.R. Nature Biotech 22, 909-911 (2004)

Best possible hairpin

- black part is given
  - what are the possibilities for *i* and *j* ?



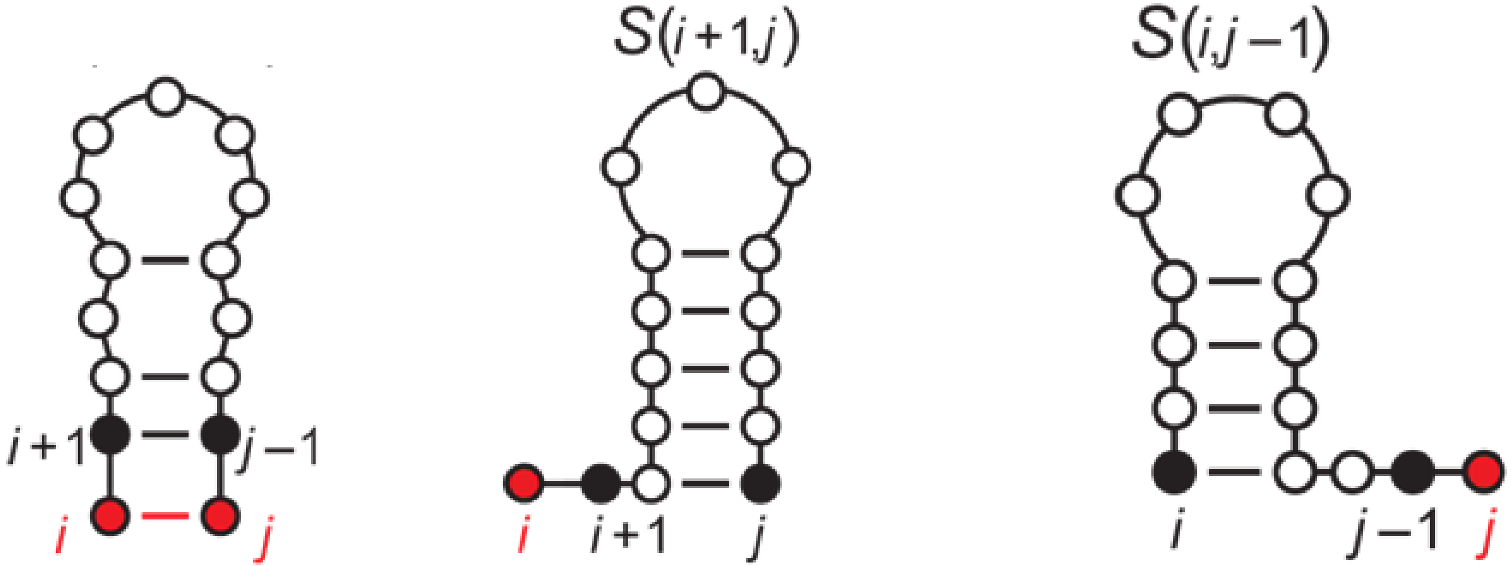
- maybe *i* should pair with *j*
- maybe there is a better *j* later

- what possibilities must one consider ?

Eddy, S.R. Nature Biotech 22, 909-911 (2004)

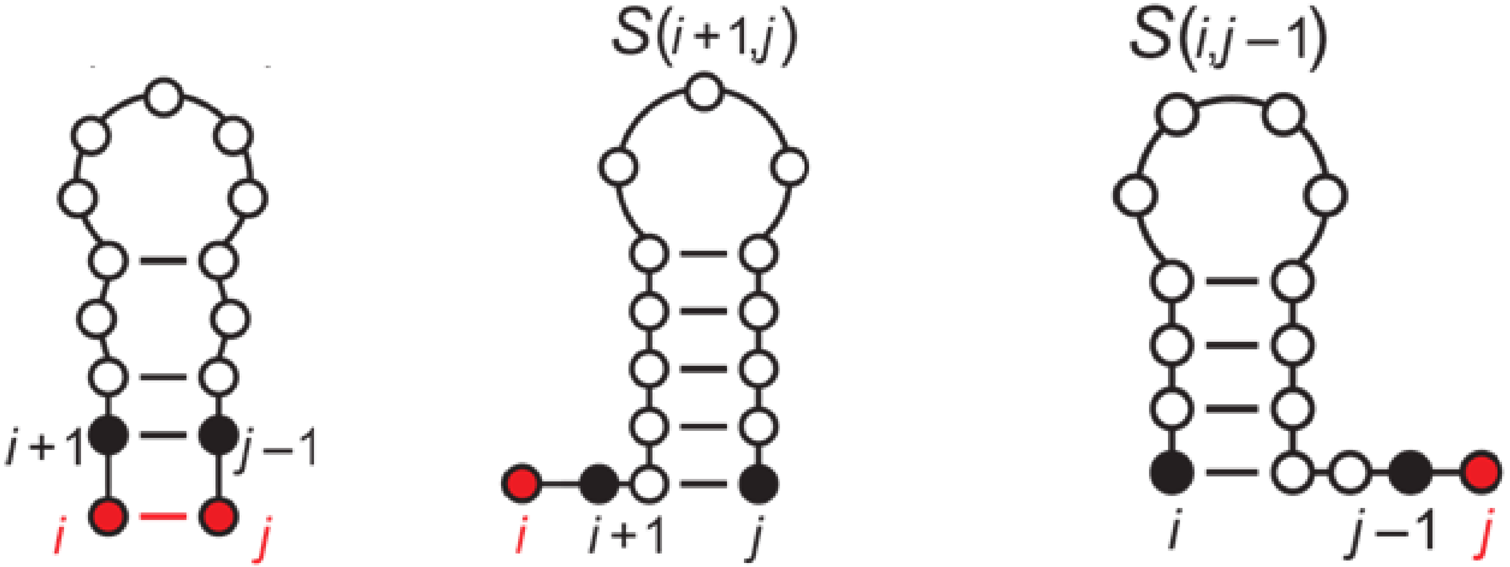
Optimal hairpins

- extend the hairpin
- put a gap / bulge in the left
- put a gap / bulge on the right



Eddy, S.R. Nature Biotech 22, 909-911 (2004)

Optimal hairpins

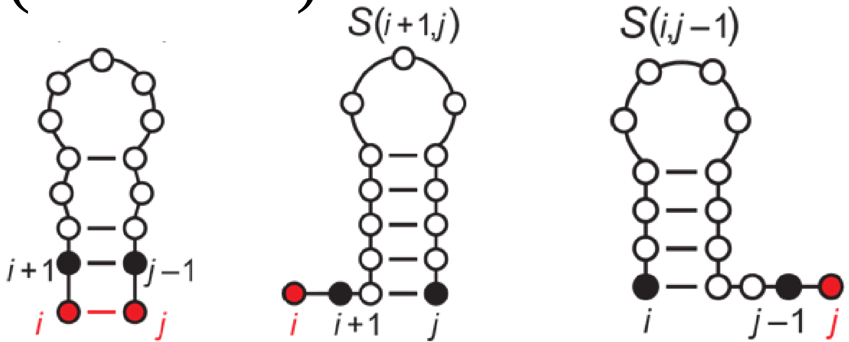


- order of steps
    - start by finding best local loops/pairs
    - move outwards

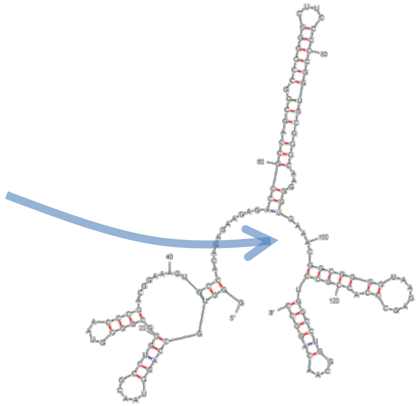
- consequence
    - base pairs will never cross - important

Optimal hairpins

- How expensive ?
  - look at all *i* positions (*n* of them)
    - look at all *j* neighbours (*n* of them)
  - O(n^2) - not finished yet



- What have we done ?
  - best organisation of hairpins
    - with best position of bulges and gaps
- Cannot yet split a chain into multiple hairpins

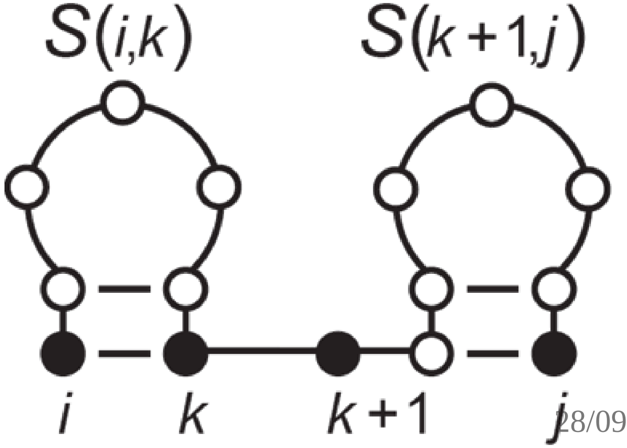


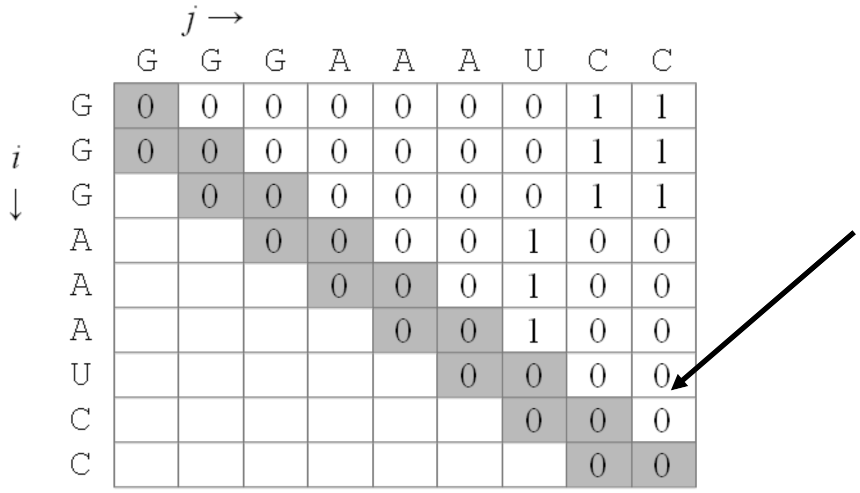
Splitting hairpins

- Check every position *k*
  - split and check the hairpin to left and right
  - check the score with every value of k

- result ?
  - for each possible position see if a split / bifurcation helps
  - at each position we have best possible hairpin
- final result ?
  - best possible set of base pairs

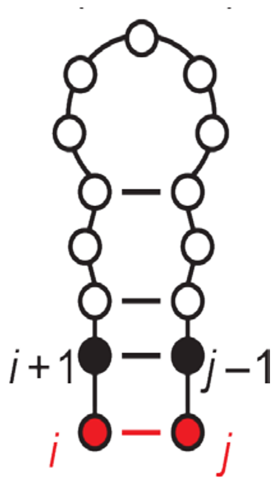
- how to implement ?





$i \downarrow$ / $j \rightarrow$	G	G	G	A	A	A	U	C	C
G	0	0	0	0	0	0	0	1	1
G	0	0	0	0	0	0	0	1	1
G		0	0	0	0	0	0	1	1
A			0	0	0	0	1	0	0
A				0	0	0	1	0	0
A					0	0	1	0	0
U						0	0	0	0
C							0	0	0
C								0	0

start here



- For each cell on diagonal,

$$S(i,j) = S(i,j)_0 + \max \begin{cases} S(i+1, j-1) \\ S(i+1, j) \\ S(i, j-1) \\ \max_{i<k<j} S(i,k) + S(k+1, j) \end{cases}$$

Scoring

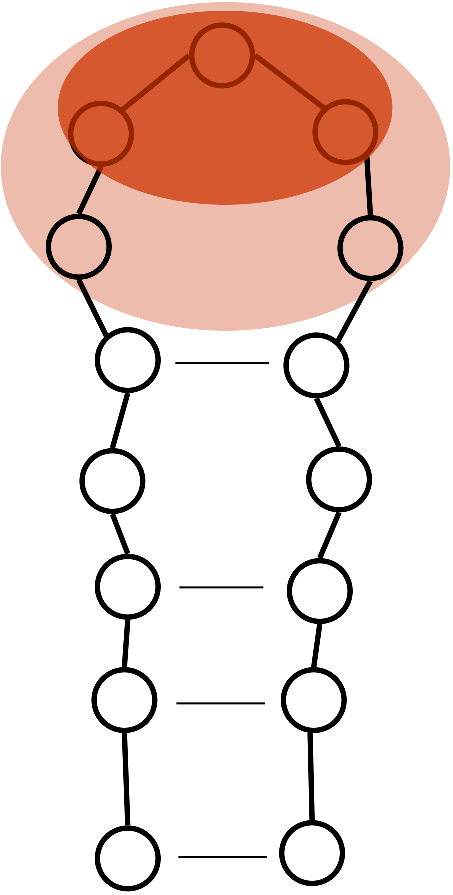
- Hydrogen bonds are good
  - GC 3 H-bonds
  - AU 2 H-bonds
  - GU 2 H-bonds
- still very crude
  - are base pairs really independent ?

- … "individual nearest neighbour model" (Matthews / Turner model)

loops / unpaired bases

- still very crude
  - loops / unpaired bases
  - counted for zero before
  - compare loop of 3 / 5 / ..
- do these bases
  - interact with each other ? solvent ?
  - energy is definitely ≠ 0

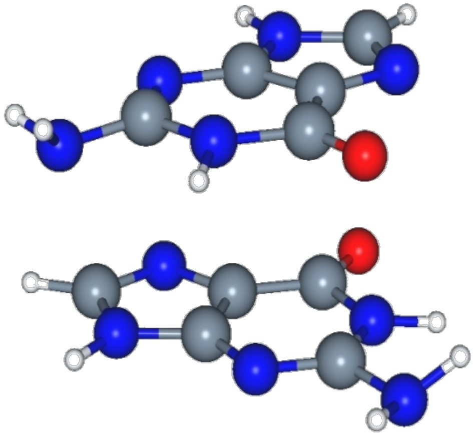
- are base pairs really independent ?
  - … "individual nearest neighbour model"



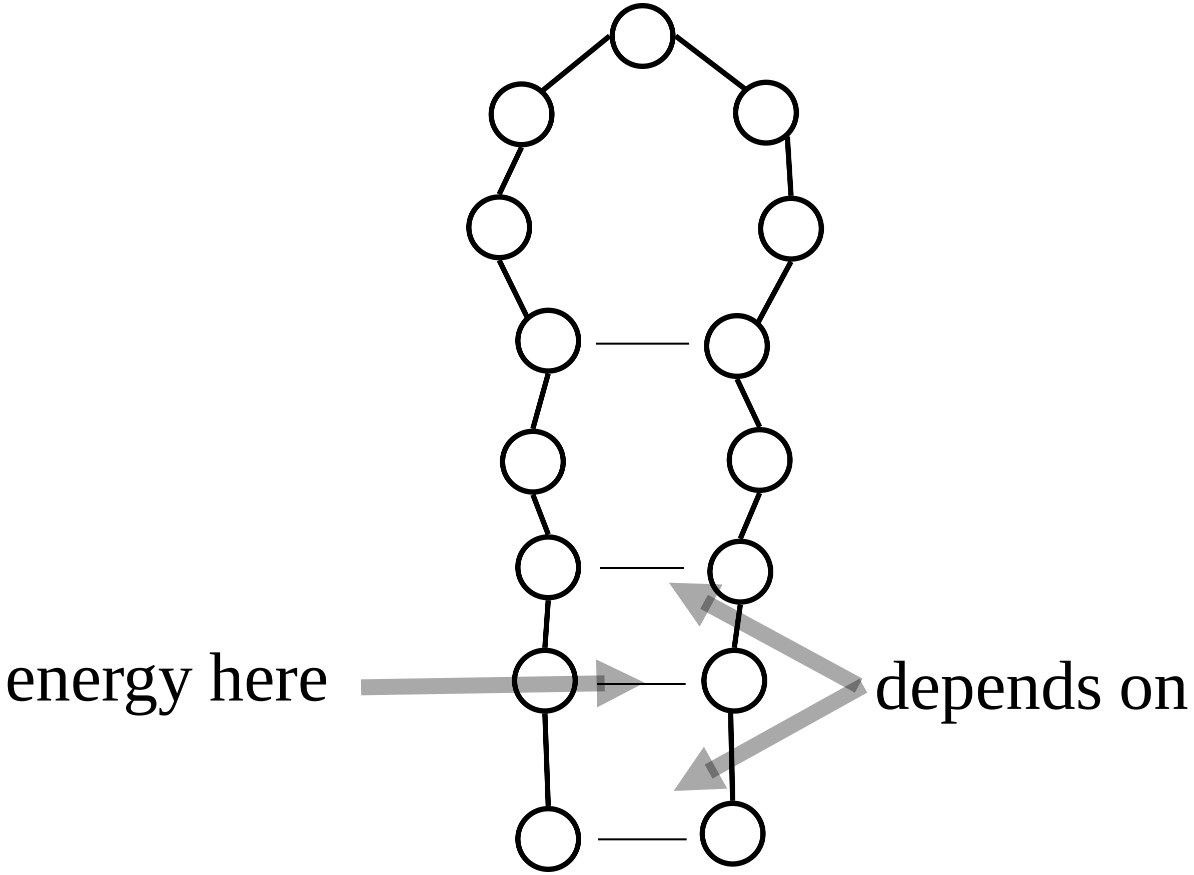
base stacking

- Originally assume base pairs are independent
  - score = sum of base pairs

  - valid ?
  - consider all the interacting planes
    - partial charges, van der Waals surfaces



Nearest neighbour model



- goal
  - incorporate most important effects
  - do not add too many parameters

Nearest neighbour model

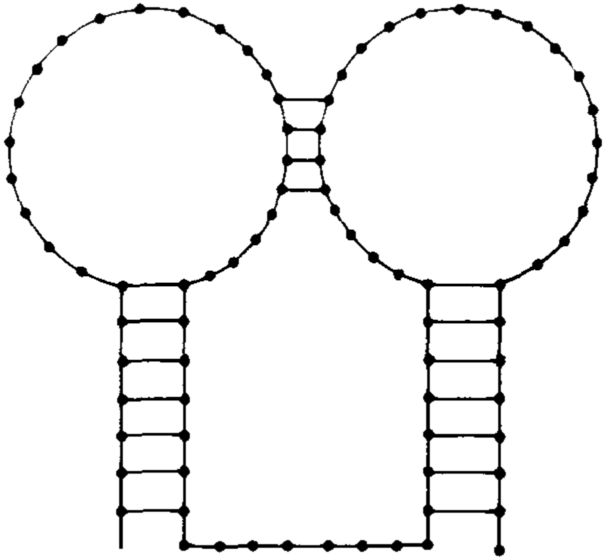
- many many parameters
- empirical
- how good ?
  - overall prediction ≈ 70 %

- problems
  - energy model fundamentally broken
  - ΔG is not pair-wise additive
  - no accounting for longer range interactions
  - worse…
    - pseudoknots

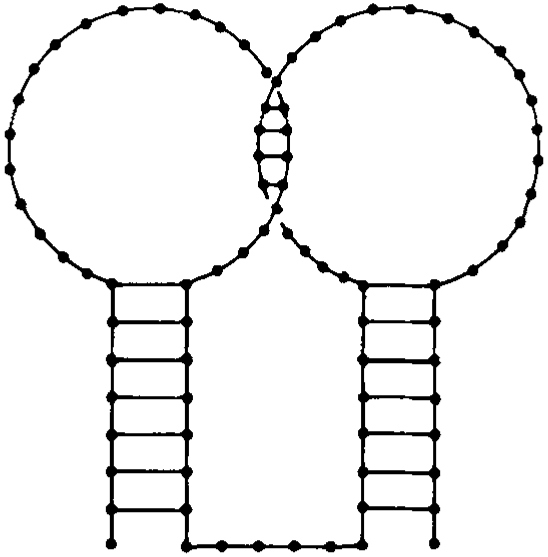
Knots

pseudo knot

- not a knot at all



real knot

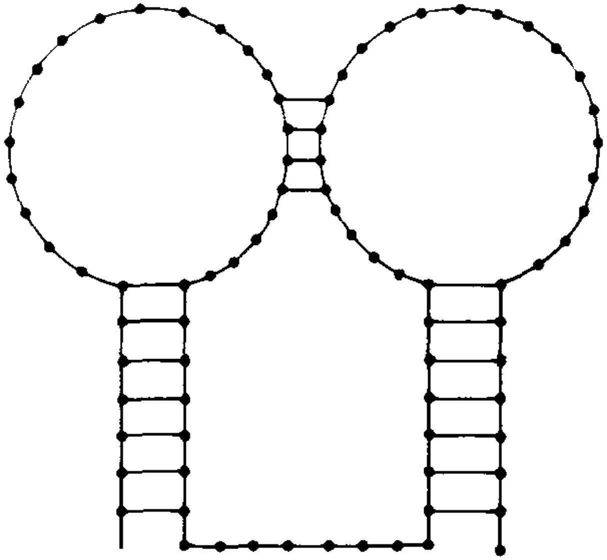


H-bond pattern is identical

- in the representations we have
  - reasonable patterns look like knots

picture from Zuker & Sankoff, Bull. Math. Biol. 4, 591-621 (1984), RNA secondary structures and their prediction

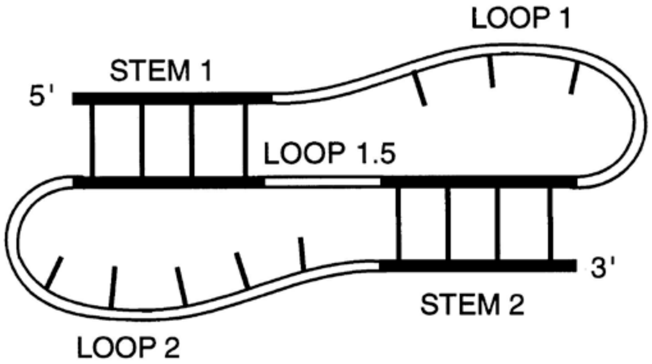
pseudoknots

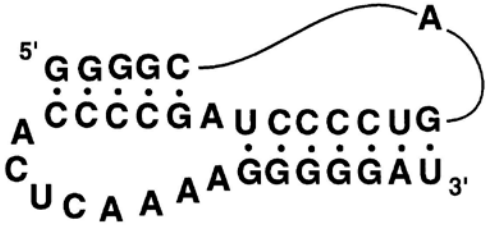


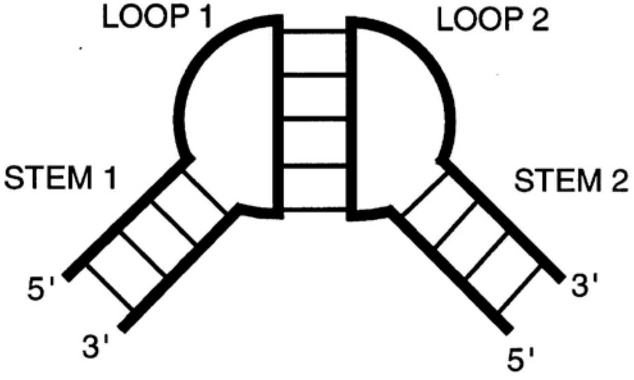
- look at pattern of H bonds
  - can I predict optimal behaviour of *i*, *j* given previous structure ?
- No !!
- Simple friendly pattern cannot be predicted
- (*i* < *i*' < *j*< *j*')

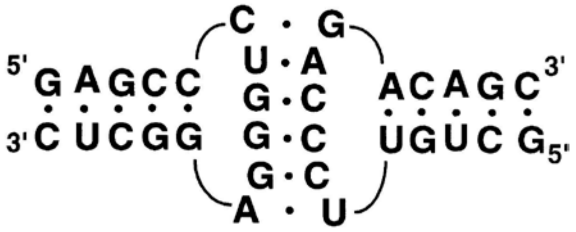
- different kinds / topologies ?

Pseudo knots

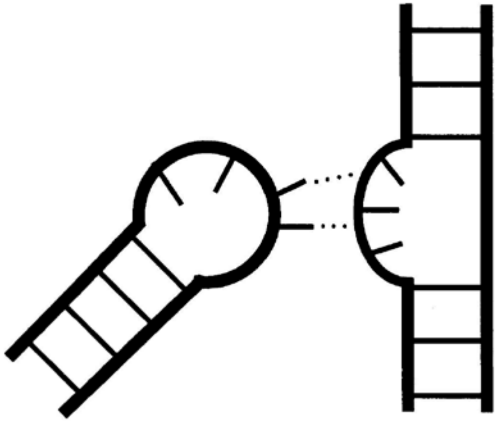


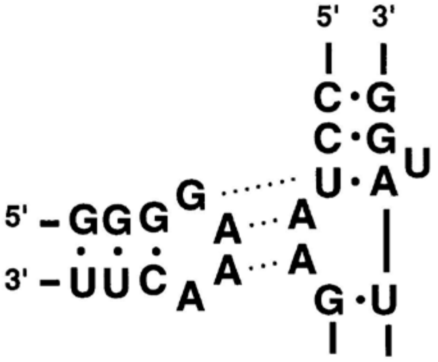






kissing hairpins





hairpin loop - bulge

from Burkard, M.E., Turner, D.H., Tinoco Jr., I., in The RNA World, 2nd Edn, eds Gesteland, RF, Atkins, JF Cold Spring Harbor Laboratory Press (1998)

Summarise

- Simple prediction O(n^3)
- with few pseudoknot types O(n^4)
- general case much worse

Active areas

- RNA interacts with proteins – prediction of these regions
- treating pseudo knots
- using related RNA's to improve reliability
- sequence design
- folding simulations
- comparison of molecules

Problems

- predictions far from reliable

- other approaches
  - non-dynamic programming ?
    - reveal problems in score functions

- only base-pair interactions considered
- everything is 2D

- kinetics ?