RNA structure, predictions

Themes

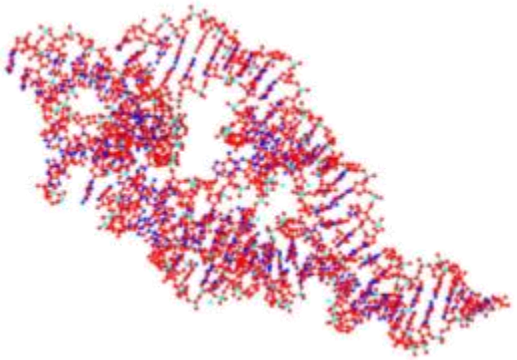
- RNA structure
  - 2D, 3D
  - structure predictions
  - energies
  - kinetics

Werfen Sie das alte Handout "RNA structure, prediction" weg !! Heute – ein viel kürzeres

Structure – protein vs RNA

Middle of proteins

- hydrophobic core
  - soup of insoluble side chains



Middle of RNA

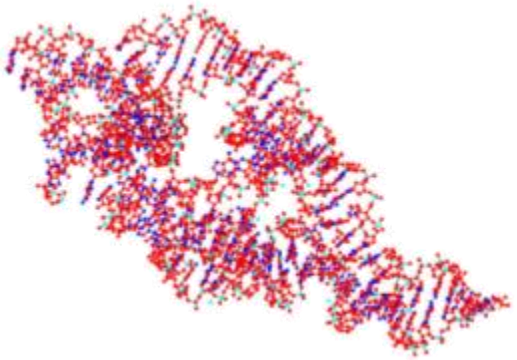
- base-pairing / H-bonds
- much more soluble
  - if something wants to forms H-bonds, there is competition from water

Protein structure lectures are not helpful today

RNA – how important is 3D structure ?

Binding of ligands (riboswitches, ribozymes)

- totally dependent on 3D shape -
  where functional groups are in space

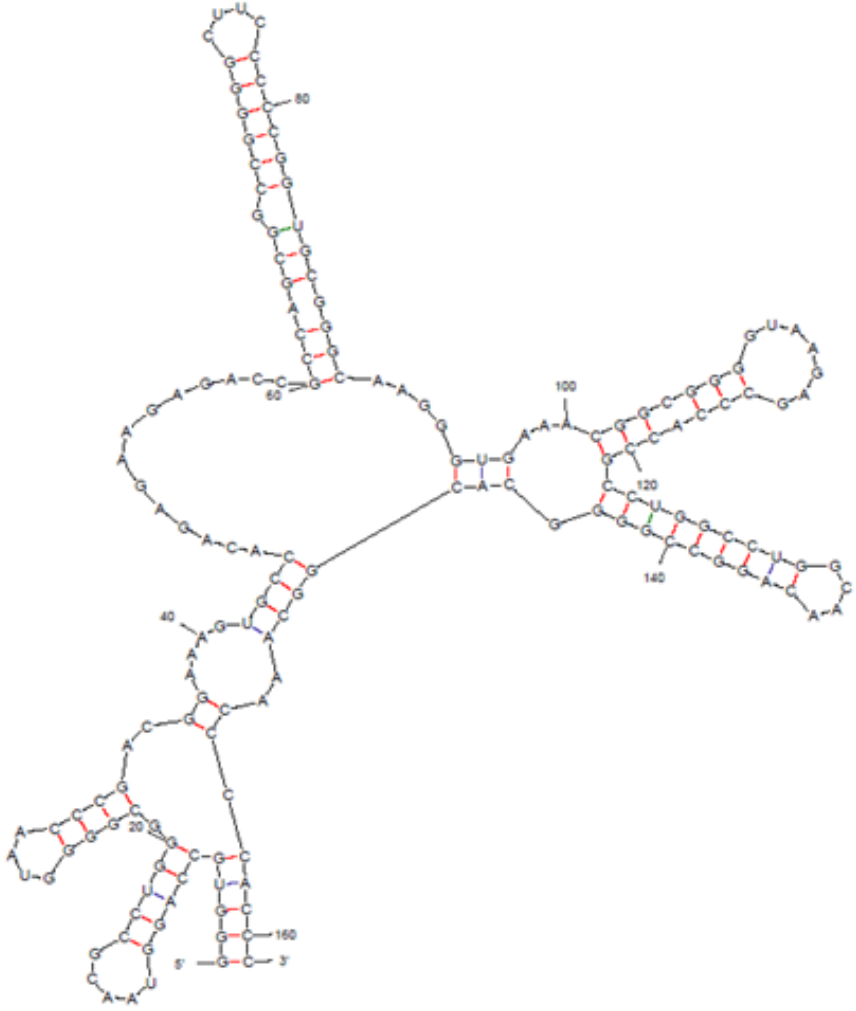


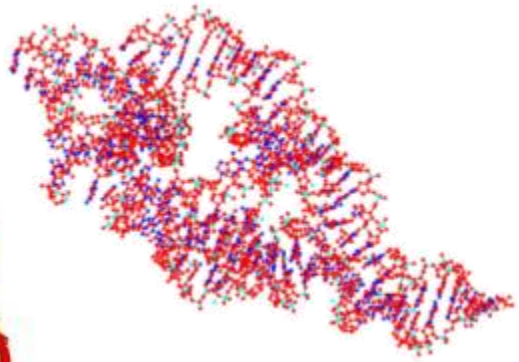
What do we do ?

- mostly ignore it

How realistic is 2D ? How relevant ?

3D versus 2D



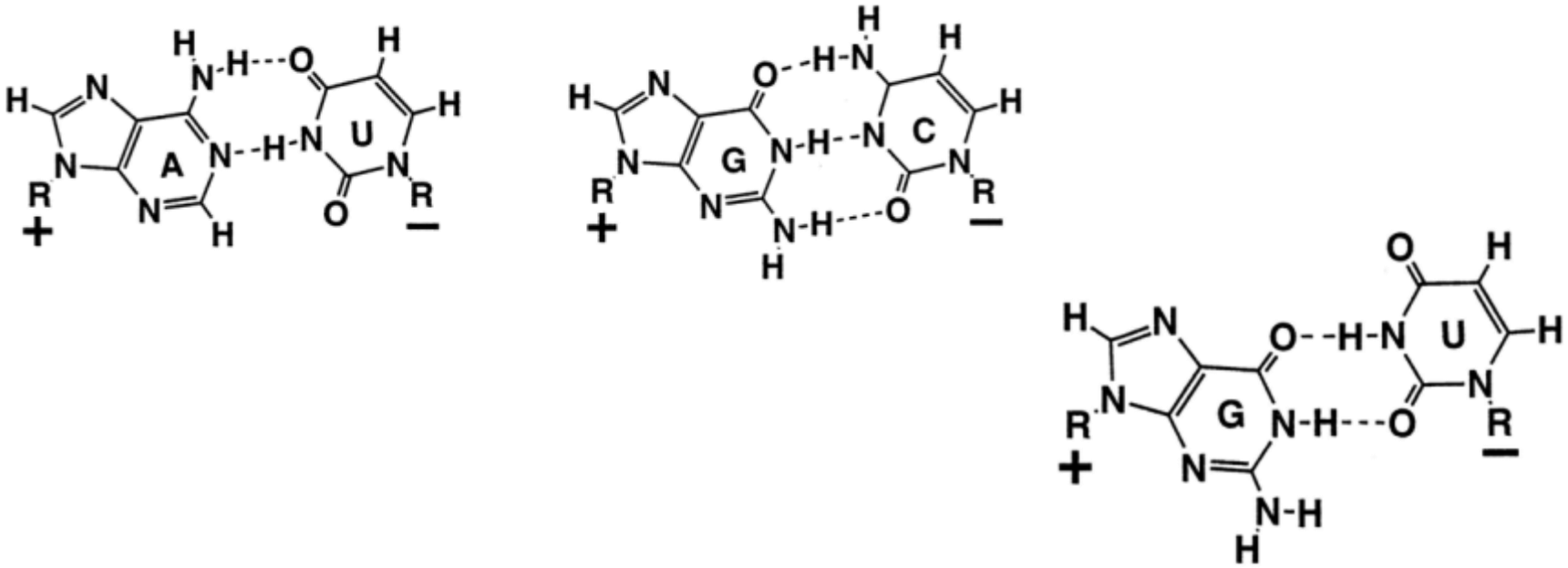


PDB acquisition code 1u9s

2D why of interest ?

1. computationally tractable (fügsam / machbar)

2. historic – belief that nucleotides are
   dominated by base pairs + helices (classic and wobble)

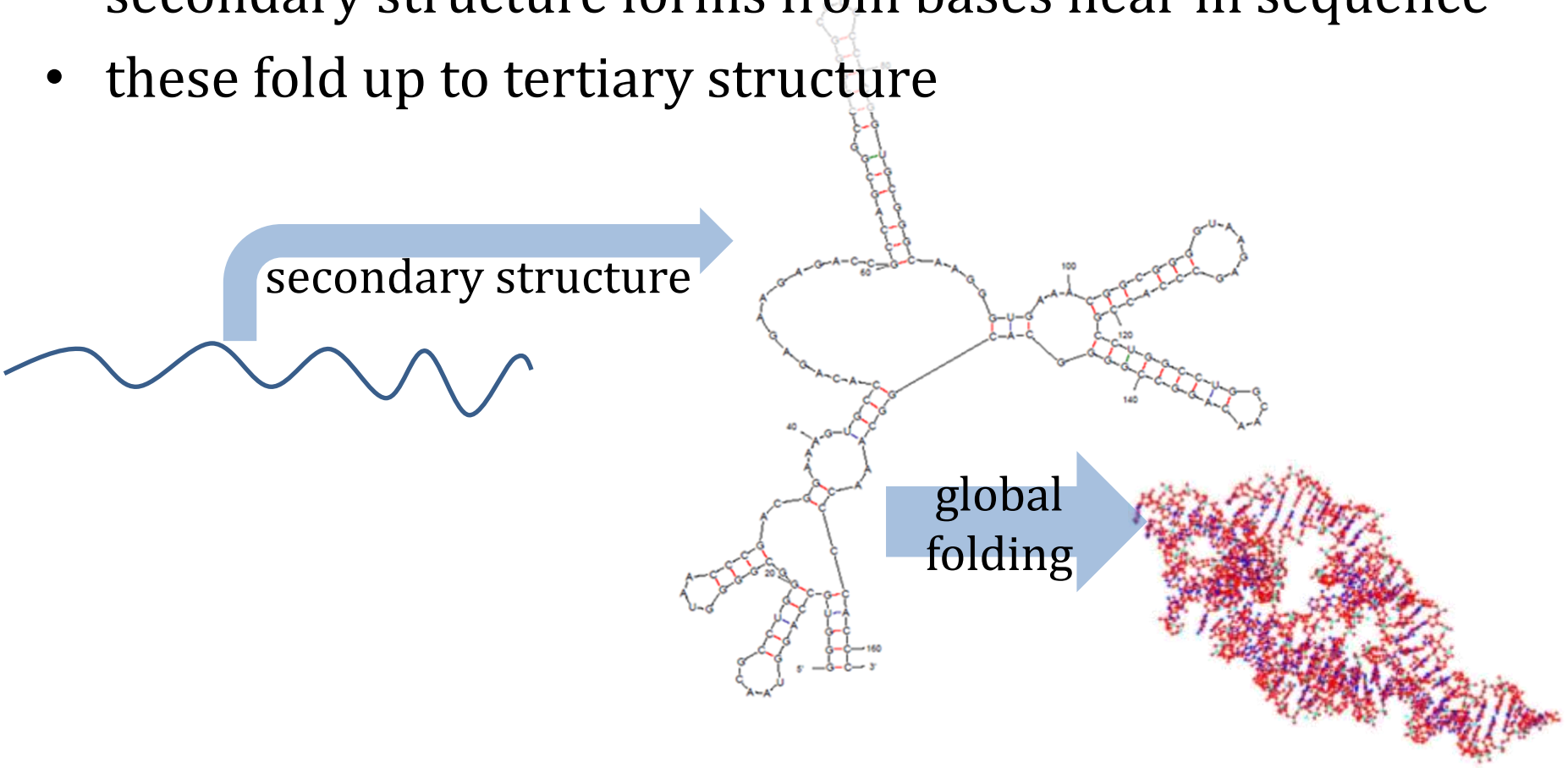


from Burkard, M.E., Turner, D.H., Tinoco Jr., I., in The RNA World, 2nd Edn, eds Gesteland, RF, Cech, TR., Atkins, JF
Cold Spring Harbor Laboratory Press (1999)

2D why of interest ?

3. Claim - RNA folds hierarchically

- secondary structure forms from bases near in sequence
- these fold up to tertiary structure



2D why of interest ?

3. Claim - RNA folds hierarchically

Contrary evidence in protein world

- isolated α-helices and β-strands are not stable in solution

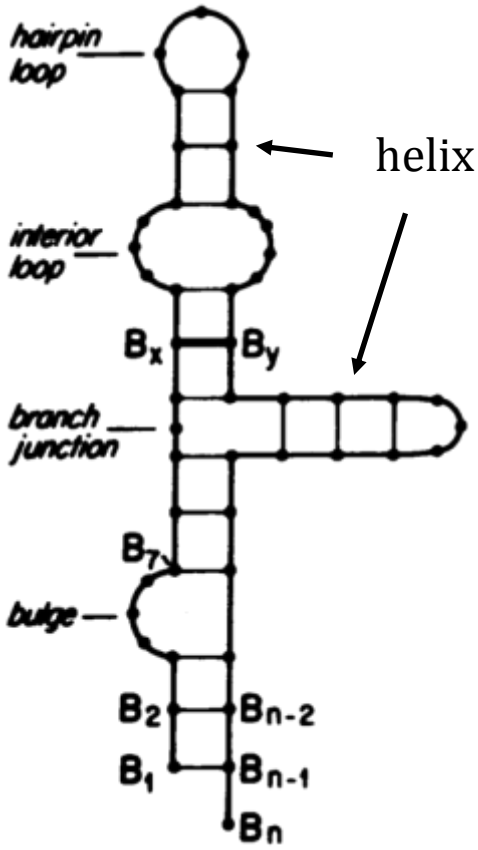
Plausible in RNA world ?

- RNA double strand helices are believed to be stable

Useful ? if true

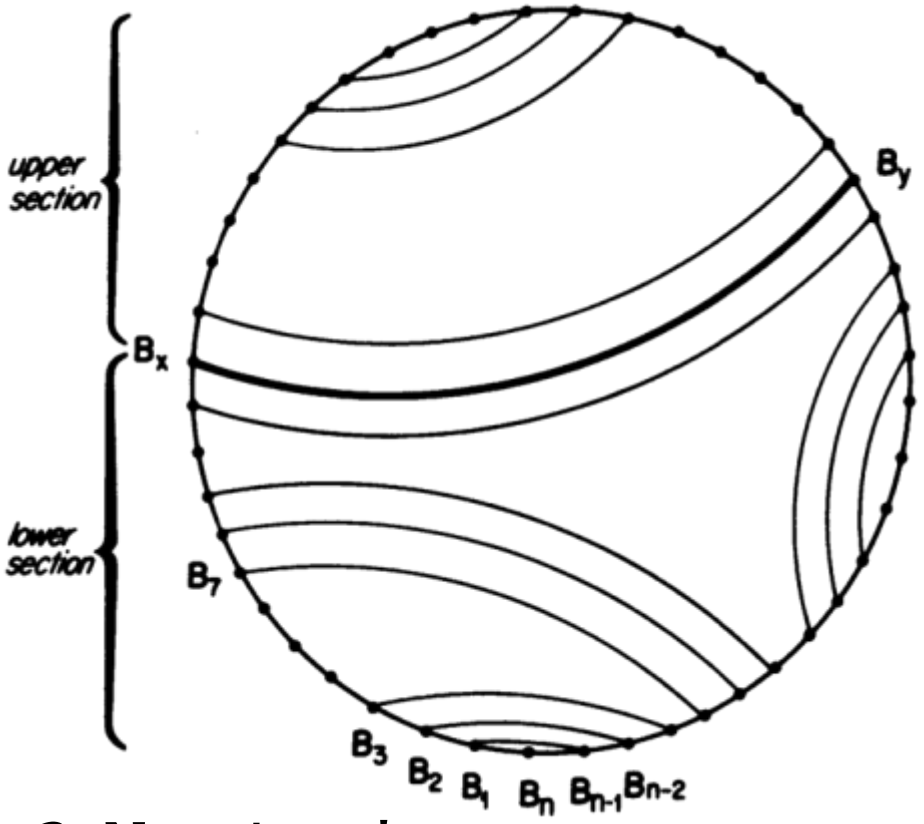
- 2D (H-bond pattern) prediction is the first step to full structure prediction

Four representations of flat RNA



1. conventional

+ on next slide

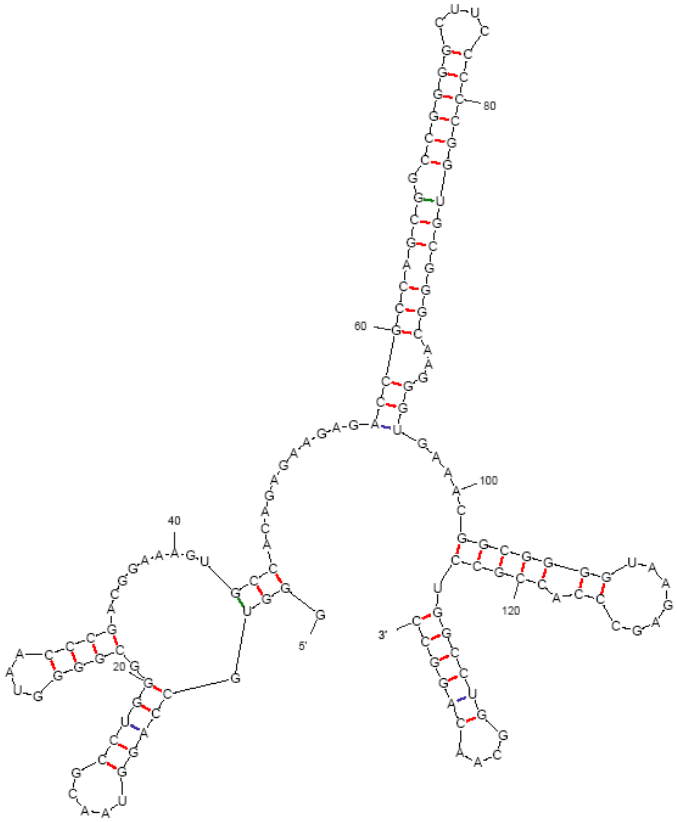


2. Nussinov's

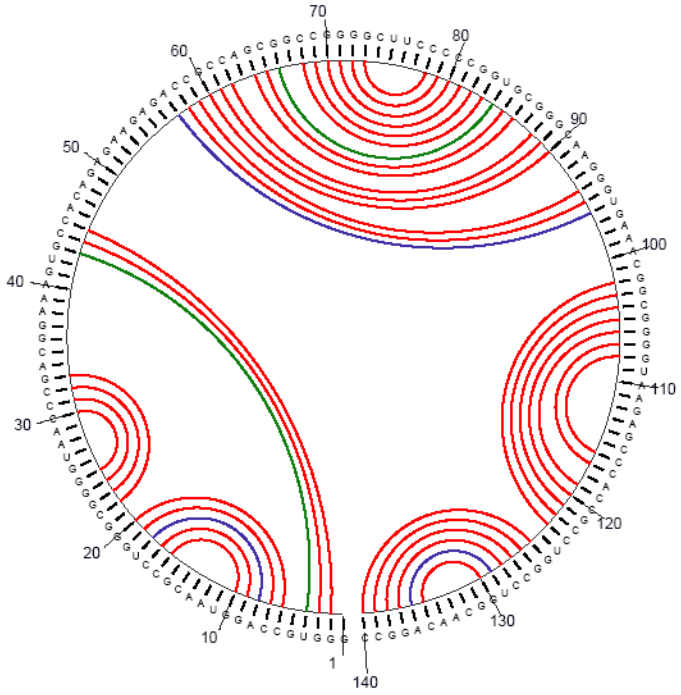
- write down bases on circle
- arcs (lines) may not cross

from Nussinov, R., Jacobson, A.B. Proc. Nati. Acad. Sci. USA, 77, 6309-6313(1980)

Four representations of flat RNA



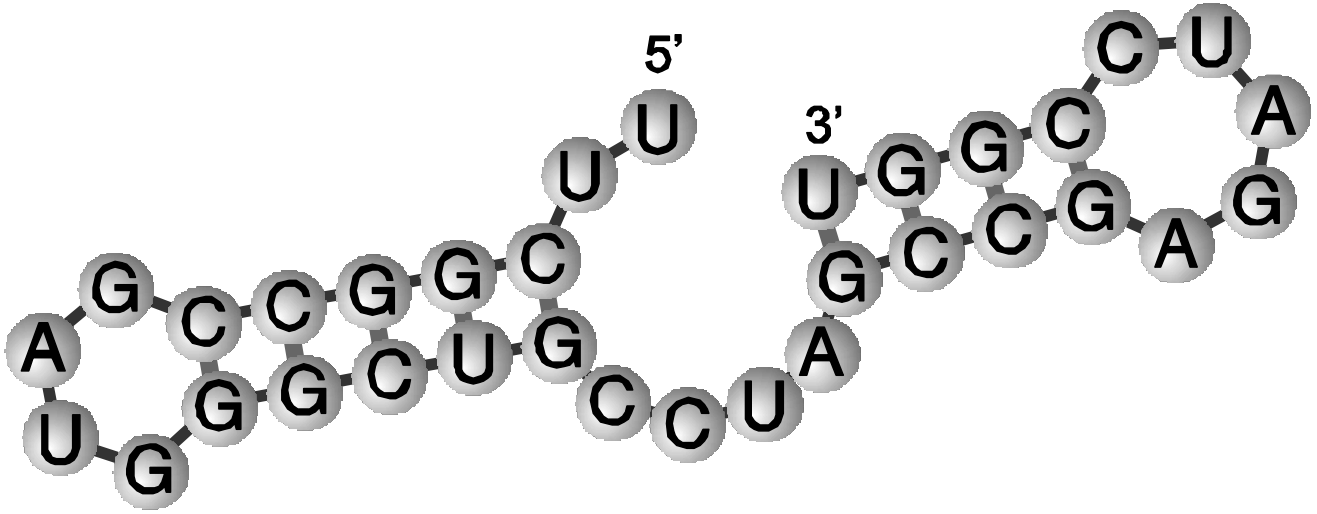
1. conventional representation



2. Nussinov's circle

Same features on both plots

Parentheses



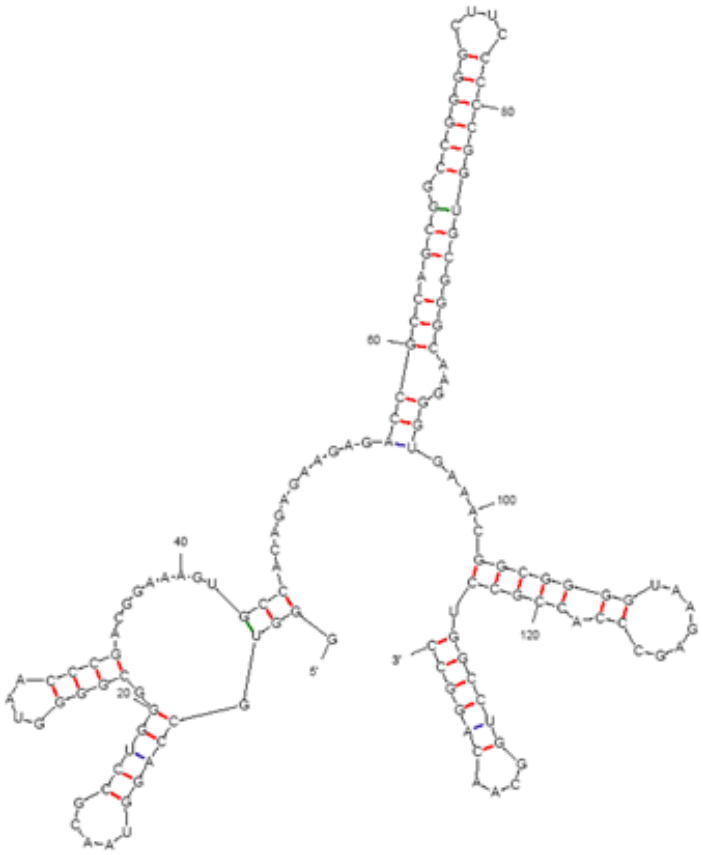
3. parentheses – most concise

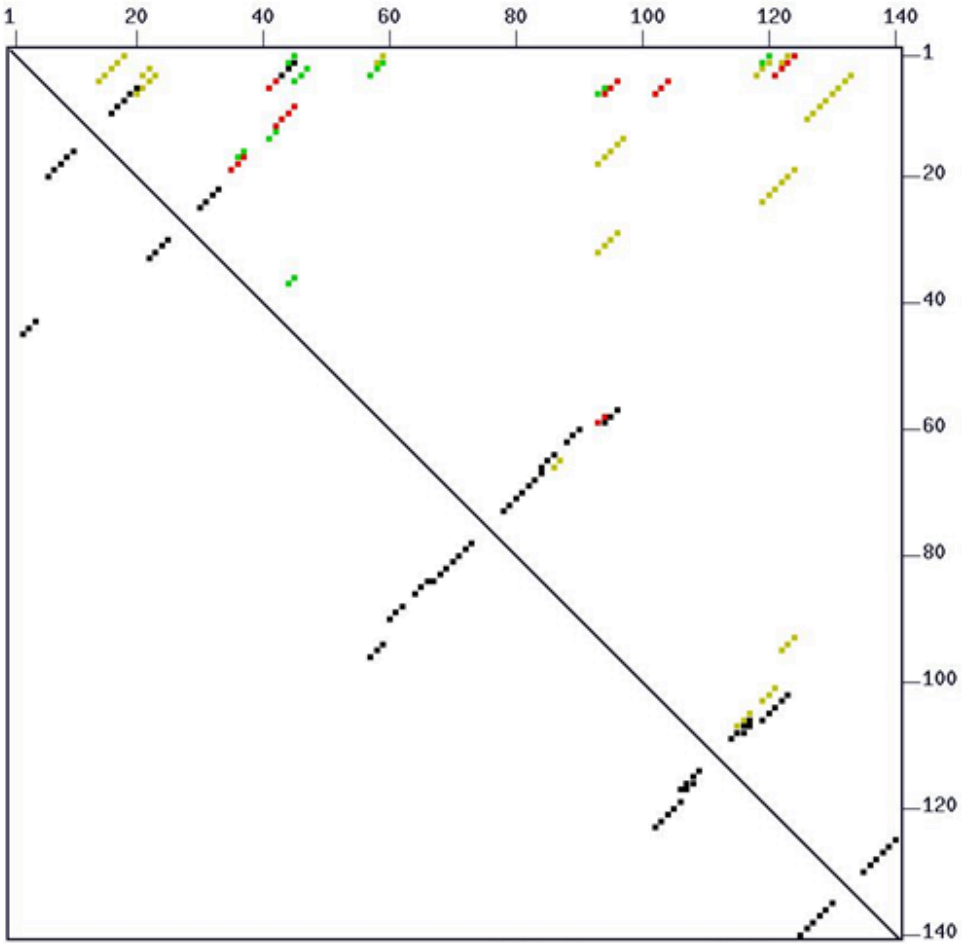
```
..(((((....)))))....(((.....)))
```

- can be directly translated to picture
- easily parsed by machine (not people)

from Schuster, P., Rep. Prog. Phys. 69 (2006) 1419–1477

Dot plots





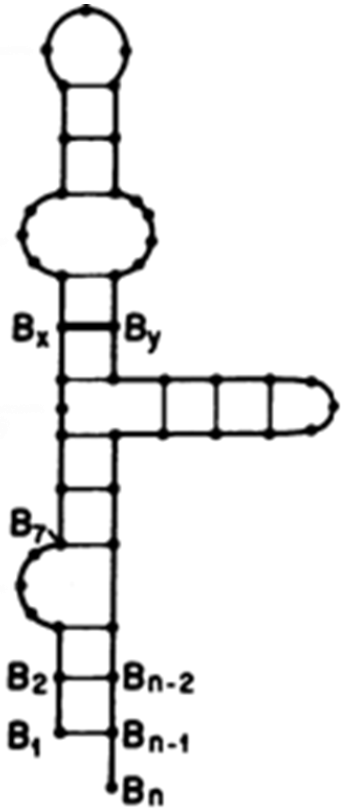
4. Dot plots

Same features in both plots

- look for long helix 57-97, bulges in long helix
- probabilities (upper right) – remember for later

made with mfold server

nomenclature / features

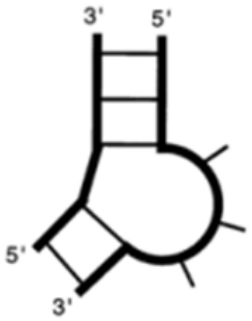


single strand

A-form double helix

Double helix with 5'-dangling end

single nucleotide bulge



three nucleotide bulge

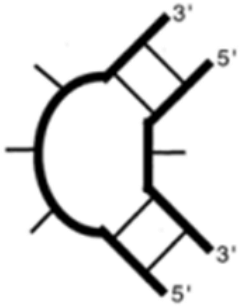
hairpin loop

For explanations later

- hairpin loop
- bulge (unpaired bases)

mismatch pair
or, symmetric internal loop of 2 nucleotides

symmetric internal loop



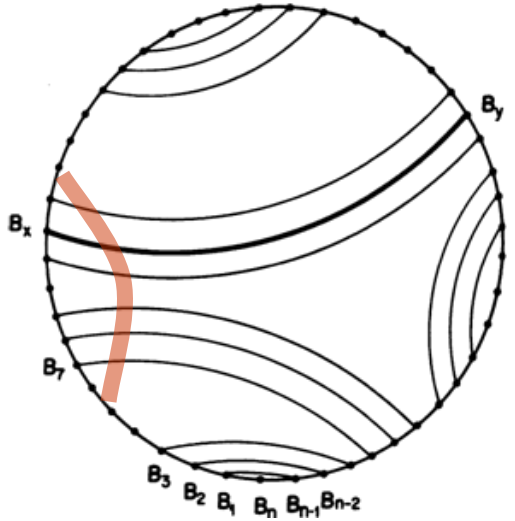
asymmetric internal loop

Nussinov, R., Jacobson, A.B. Proc. Nati. Acad. Sci. USA, 77, 6309-6313(1980)
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 (1999)

2D – properties and limitations

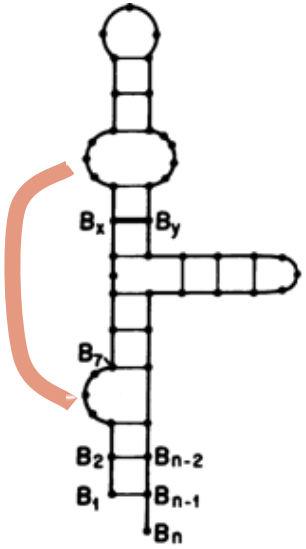
Declare crossing base pairs illegal

- think of parentheses
- discussed later



What do energies depend on ? (for now)

- just the identity of the partners
- 2 or 3 types of interaction
  - GC, AU, GU



What is the best structure for a sequence ?

Predicting secondary structure

How many structures are possible for n bases ?

$$cn^{3/2}d^n$$

for some constants c and d

- exponential growth (d^n)

Problem can be solved

- restriction on allowed structures
- clever order of possibilities

Best 2D structure (secondary)

Scoring scheme :

- each base pair scores 1 (more complicated later)

Problem

- some set of base pairs exists – maximises score

Our approach

- what happens if we consider all hairpins ?
- what happens if we allow hairpins to split in two pieces ?

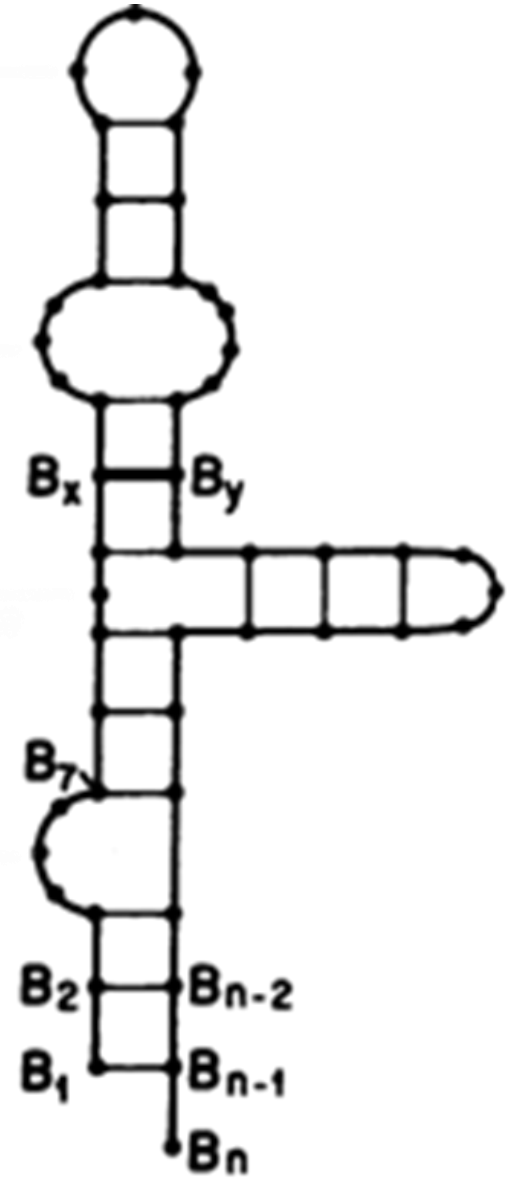
Philosophy

Structure is

- best set of hairpins (loops)
  - with bulges
    - loops within loops

Start by looking at scores one could have

- try extending each hairpin



hairpins / loops

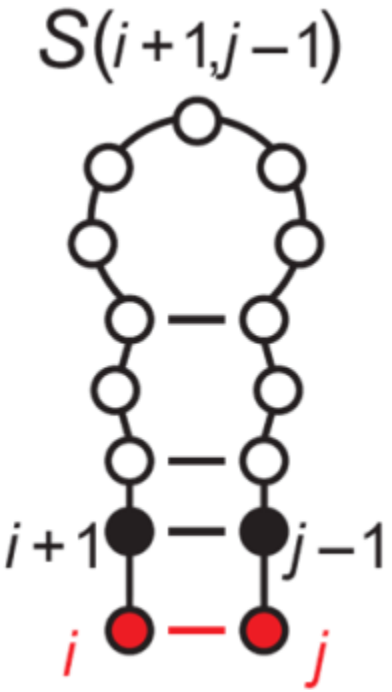
Start by looking for best possible hairpin

If we know the structure of the inner loop

- we can work out the next

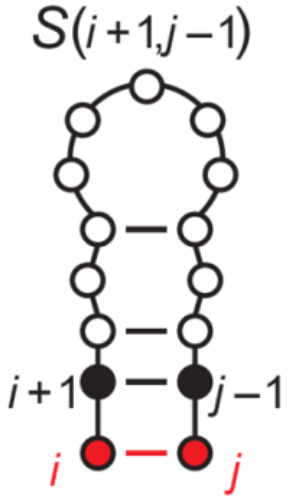
If we know the black parts

- we can decide what to do with the red ***i*** and ***j***



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

hairpins / loops

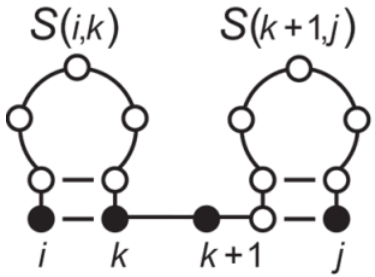


Important idea

- if I know the optimal inner loop
  try to extend it
- try to insert gaps - see if score is improved

Next important point

- walk along sequence 1..n see if score is better with two loops



Guarantees optimal solution, but…

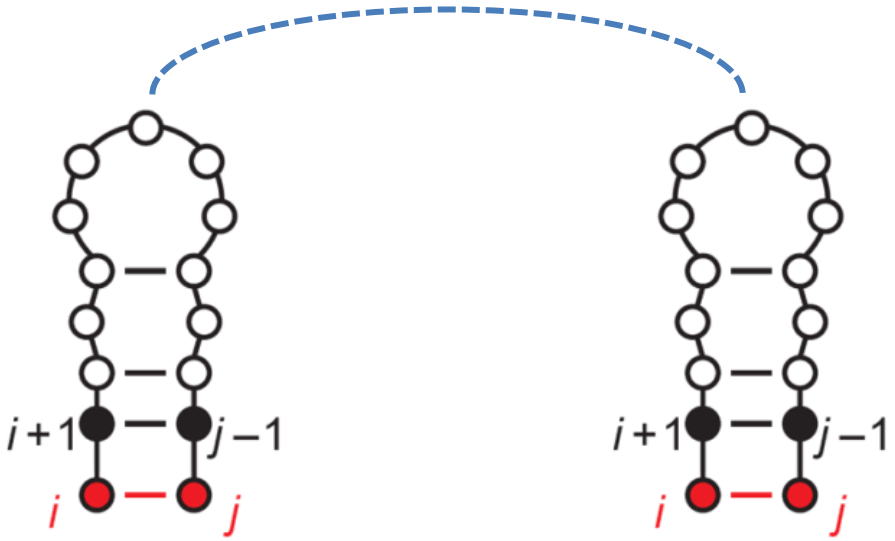
picture from Eddy, S.R. Nature Biotech 22, 909-911 (2004)

Pseudoknots

Have we considered .. ?

No !

Name – pseudoknot



Do we worry ?

- Stellingen – no
- here ? Probably.

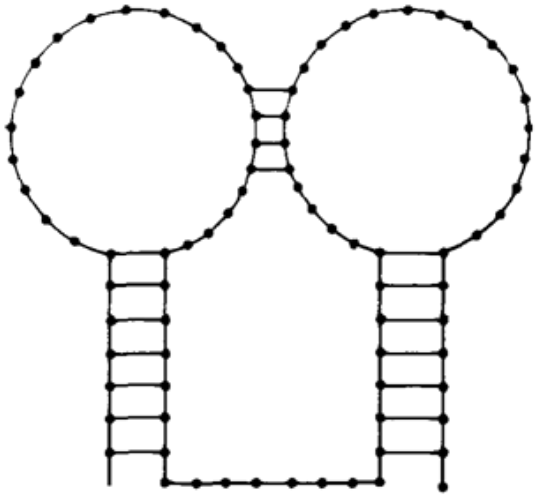
Pseudoknots

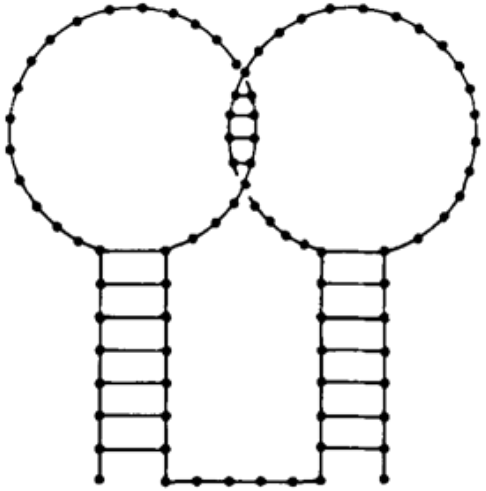
Pseudo-knot – not a knot

- why the name ?

Topologically like a knot

Would you expect them to occur ?



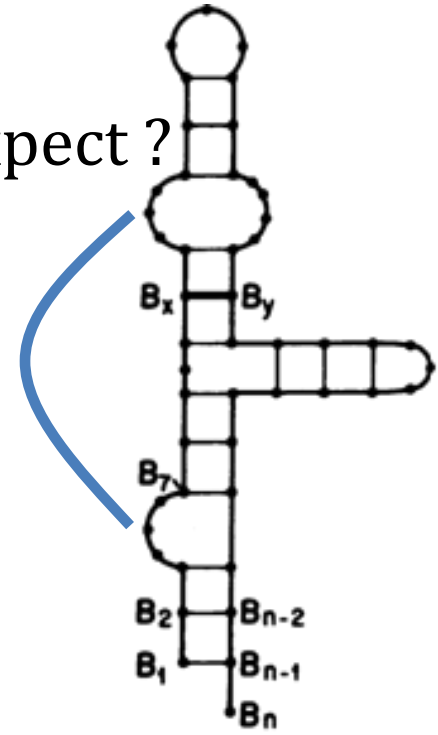


picture from Zuker & Sankoff, Bull. Math. Biol. 4, 591-621 (1984),
RNA secondary structures and their prediction

Pseudoknots

Given some unpaired bases, what would you expect ?

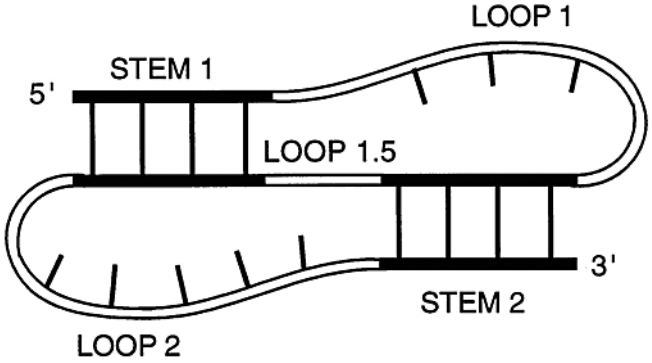
- solvate ?
- form more H-bonds ?
- pack bases against each other ?

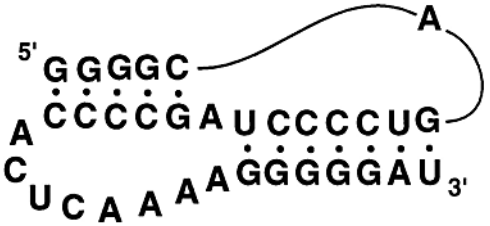


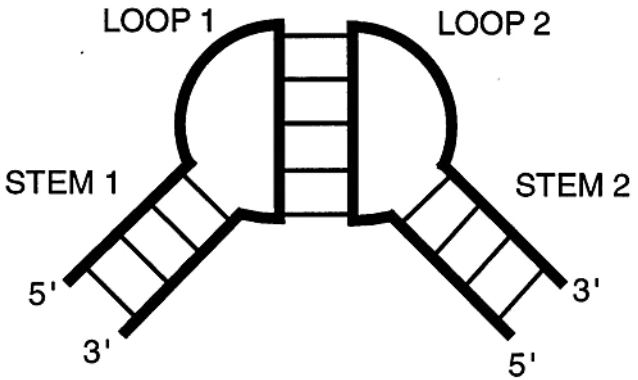
Cannot (practically) be predicted

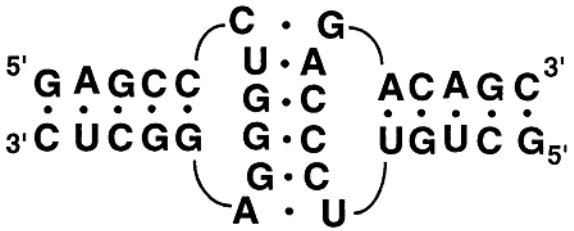
- order of steps in base-pairing methods

pseudoknots

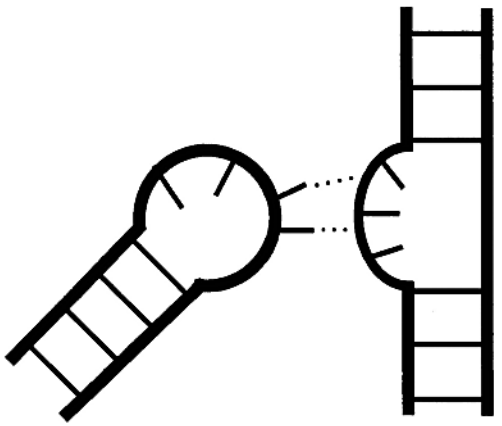


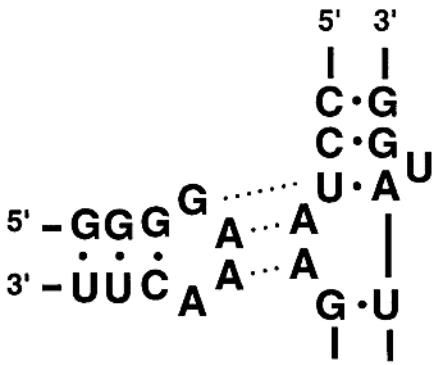






kissing hairpins





hairpin loop - bulge

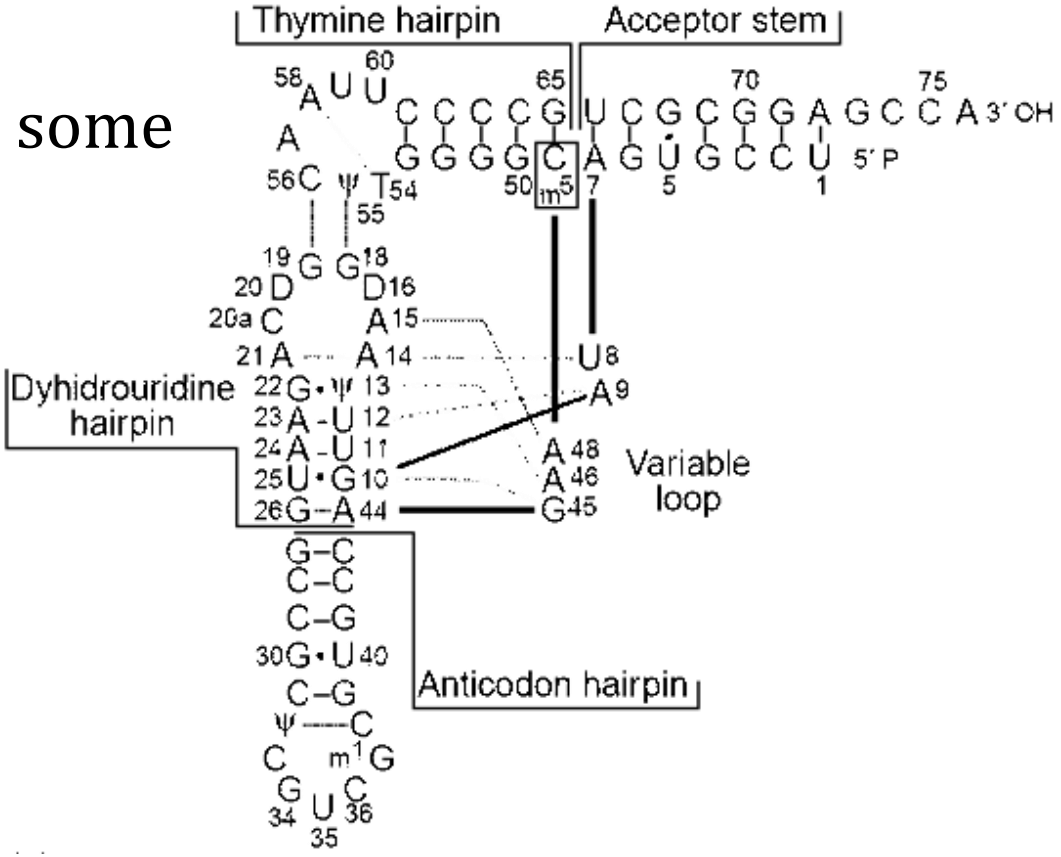
pseudoknots

Frequency of pseudoknots ?

- a few % of all H-bonds / base pairs

Significant ?

- most structures will have some
- classic RNA example



Westhof, E., Auffinger, P. in Encyclopedia of Analytical Chemistry R.A. Meyers (Ed.), John Wiley & Sons Ltd, Chichester, 2000

pseudoknot summary

Fast algorithms cannot find pseudoknots

- in order to go fast, the algorithms work in a special order
- some base pairs come in "wrong" order
- most web servers, fast programs ignore the problem

A real limitation in the methods

How expensive are the methods ?

cost of predicting structure..

The methods are not perfect.. How expensive are they ?

for each i (growing loops)

test each j

try each k (splitting loops)

gives $n \times n \times n = O(n^3)$

Scoring schemes – H bonds

Till now – count base pairs, but

We know

- GC 3 H-bonds
- AU 2 H-bonds
- GU 2 H-bonds

Compare a structure with

- 3 × GC versus 4 × AU
- 9 H-bonds versus 8 H-bonds

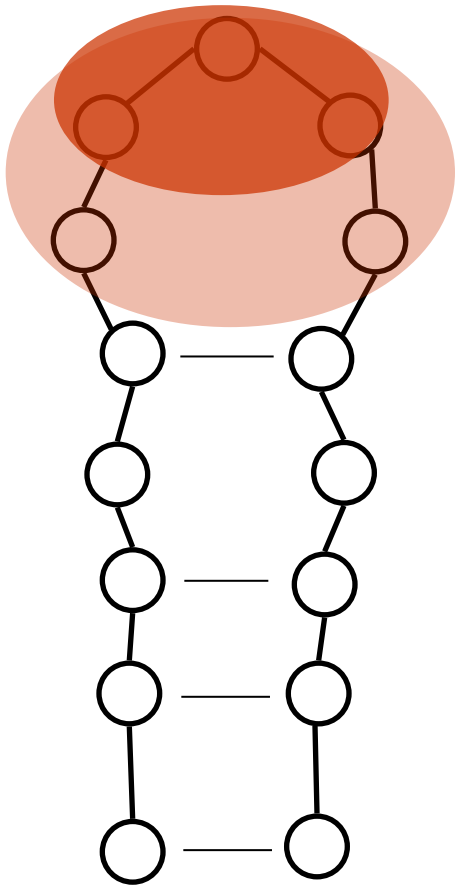
Scoring schemes – unpaired bases

Consider unpaired bases

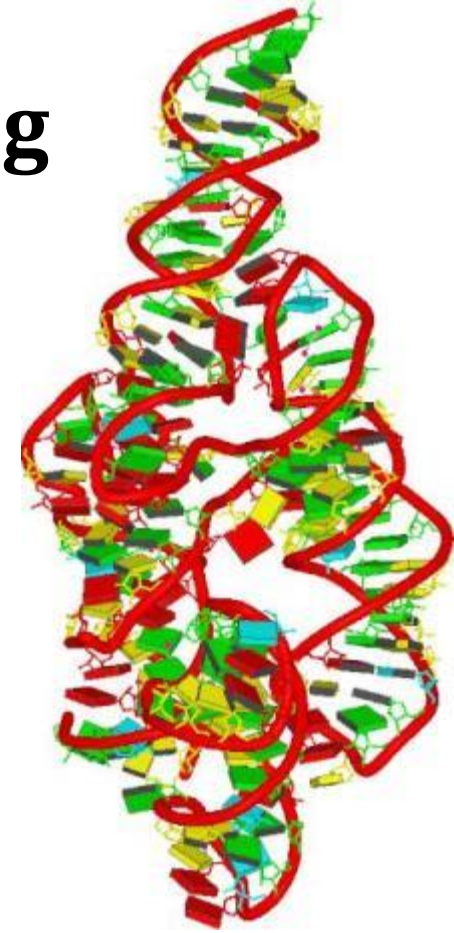
- counted for zero before
- compare loop of 3 / 5 / ..

Do these bases

- interact with each other ? solvent ?
- energy is definitely ≠ 0



Scoring schemes - stacking

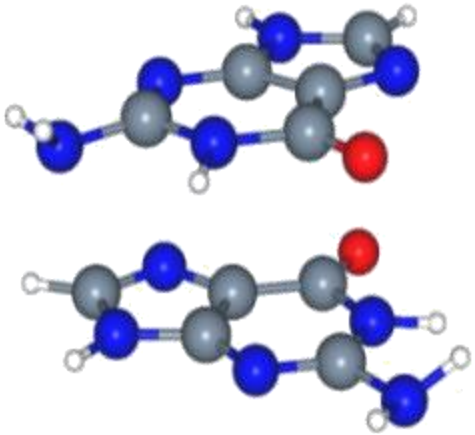


Assumption: each basepair is independent

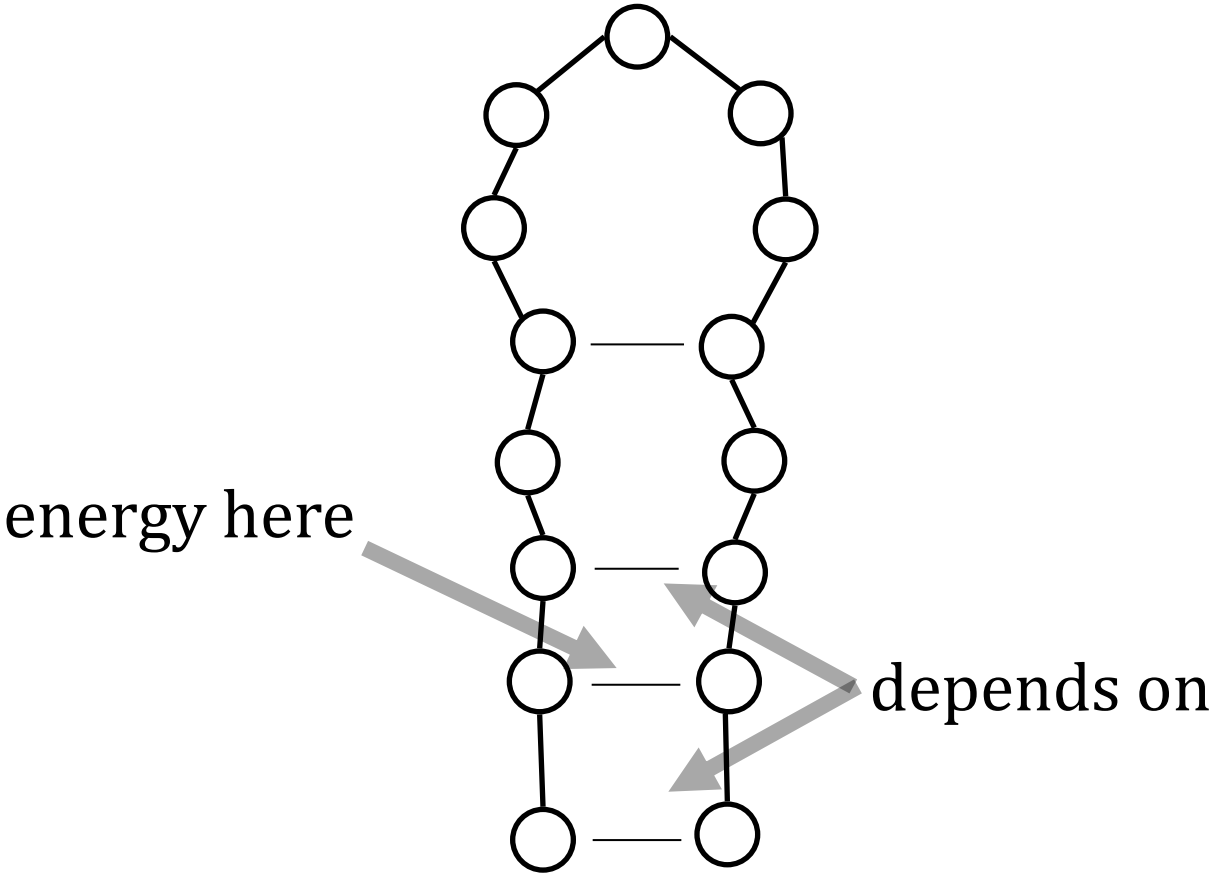
- $S(i,j)$ = base-pair + $S(i+1, j-1)$

Consider all the interacting planes

- partial charges, van der Waals surfaces



Scoring schemes - stacking



Goal

- incorporate most important effects
- do not add too many parameters … nearest neighbour model

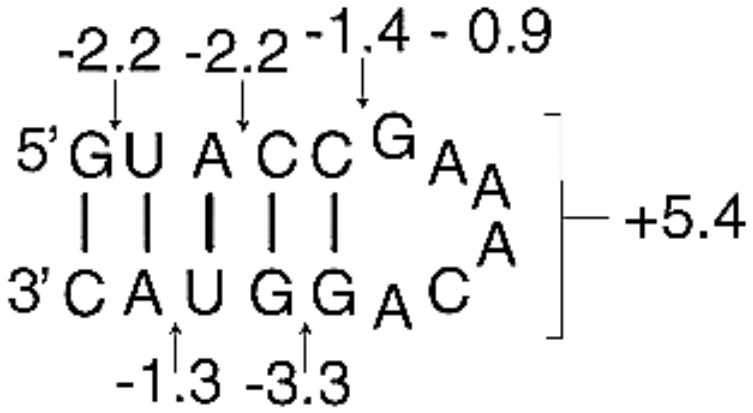
Nearest neighbour model

Previously we added

- GC + UA + AU + …

Now

- (GU/CA) + (UA/AU) +..



- terminal loop costs 5.4 kcal mol^{-1}

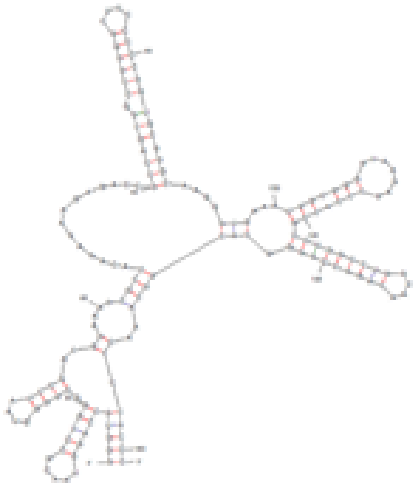
One more issue with scores and interactions..

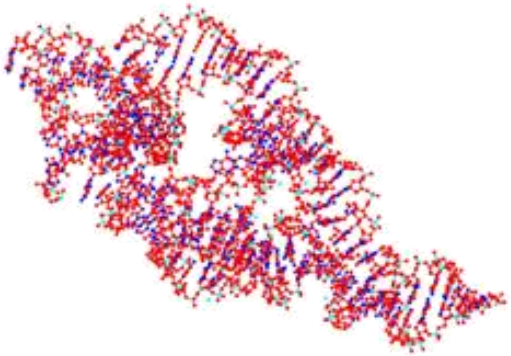
Mathews, DH, Schroeder, SJ, Turner, DH, Zuker, M in The RNA World 3rd ed, eds Gesteland, RF, Cech, RT, Atkins, JF, CSHL Press (2006)

Tertiary interactions

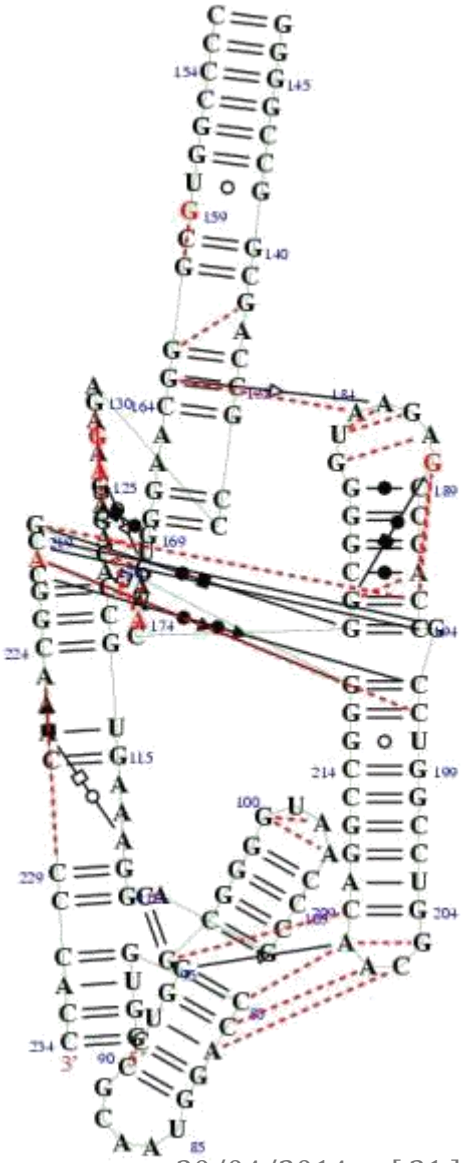
- miscellaneous H-bonds
- non-specific van der Waals

- Most larger RNA's have many tertiary interactions
- relatively compact



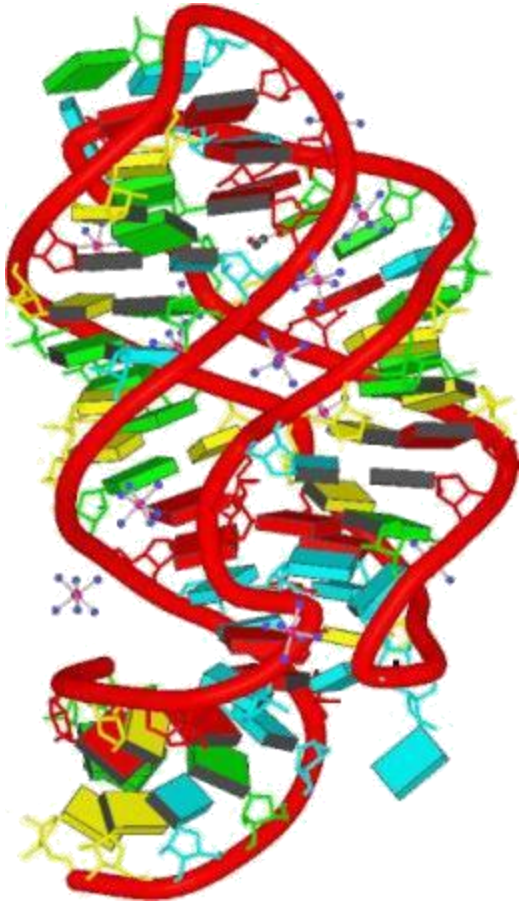


tertiary interactions from crystal

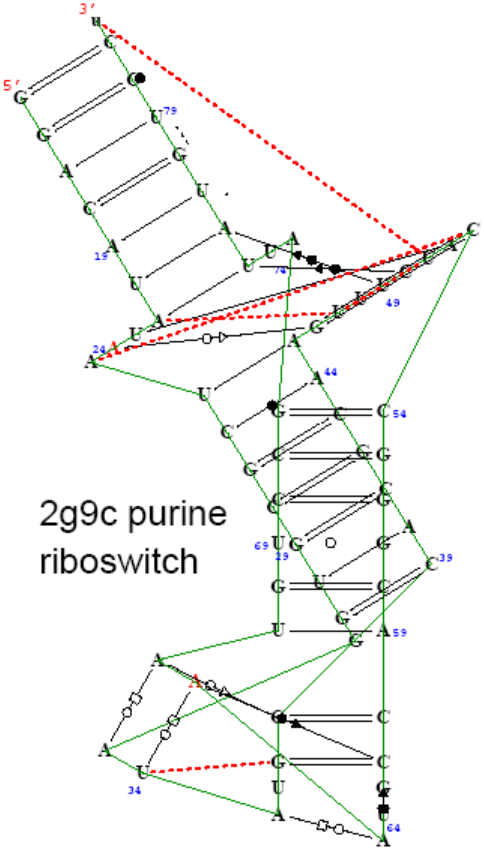


2D vs 3D

2g9c riboswitch

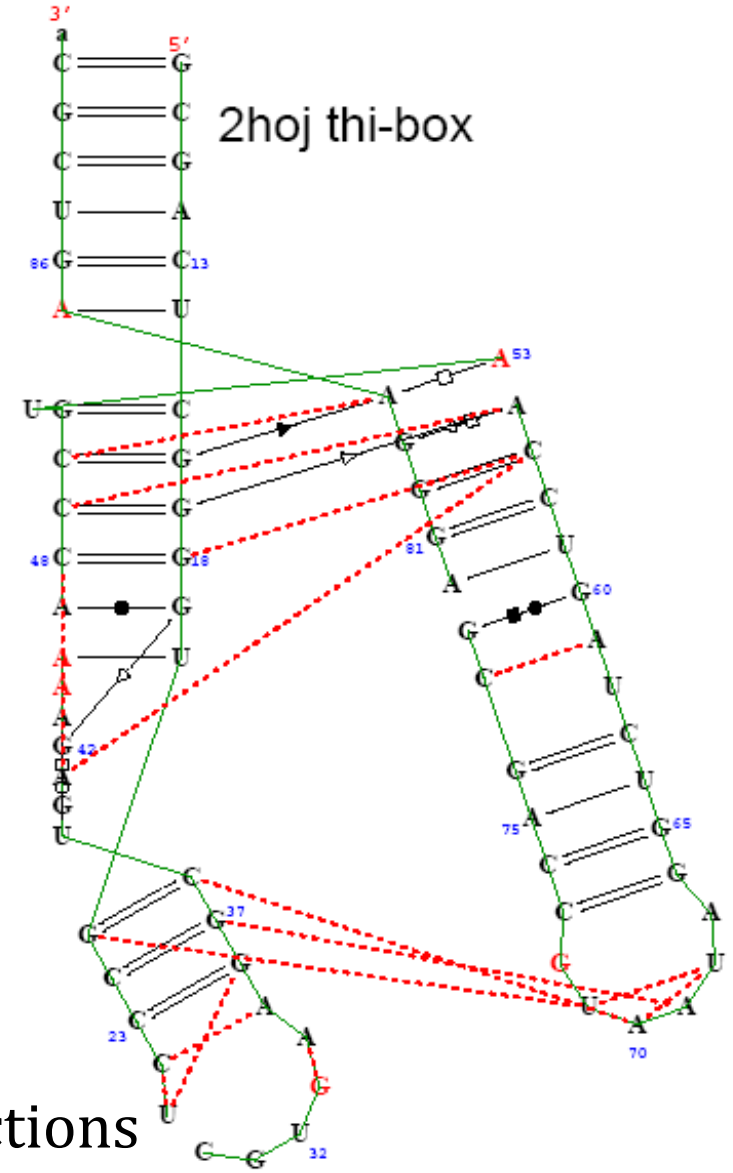


tertiary interactions from crystal



2D vs 3D

2hoj



tertiary interactions
from crystal

scoring summary

- Approximation to free energies - $\Delta G_{folding}$

n base pairs	very primitive
n H-bonds	
loop sizes	
base-stacking	nearest neighbour model
tertiary interactions	ignored

Reliability

How accurate ?

- maybe 5 – 10 % errors in energies

How good are predictions ?

- maybe 50 – 75 % of predicted base pairs are correct

Why so bad ?

Reliability

Think of an "A"

- wants to pair with a U
- there are many many U's

Think of any base

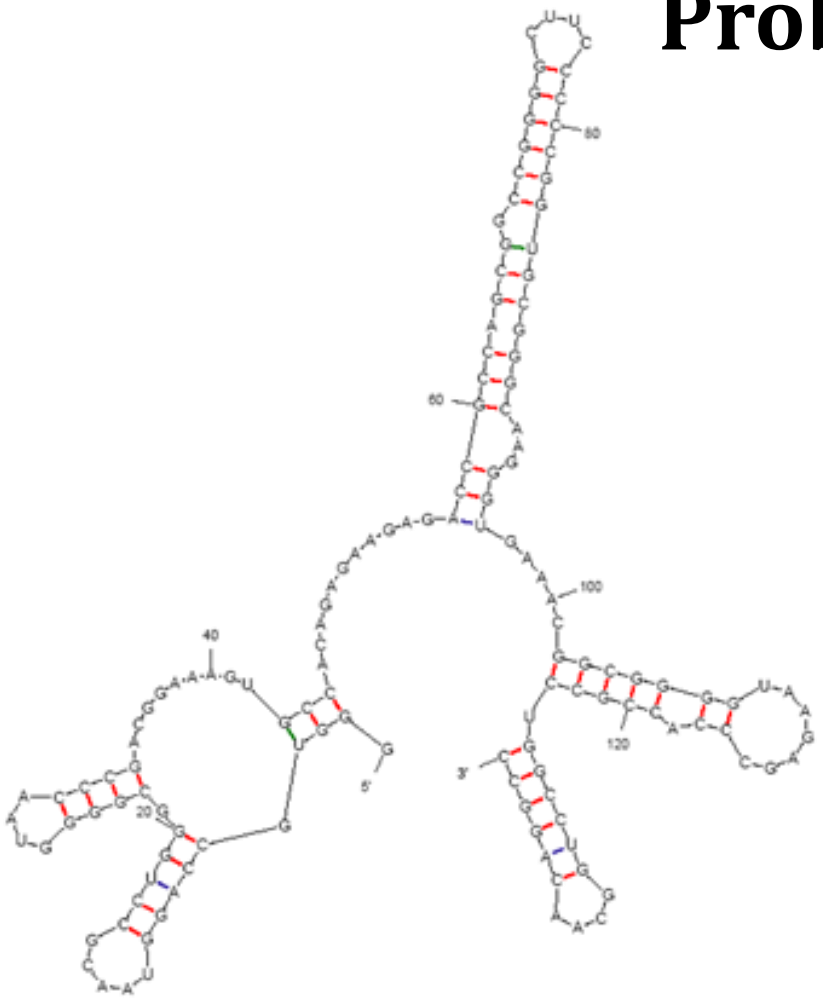
- many possible good partners

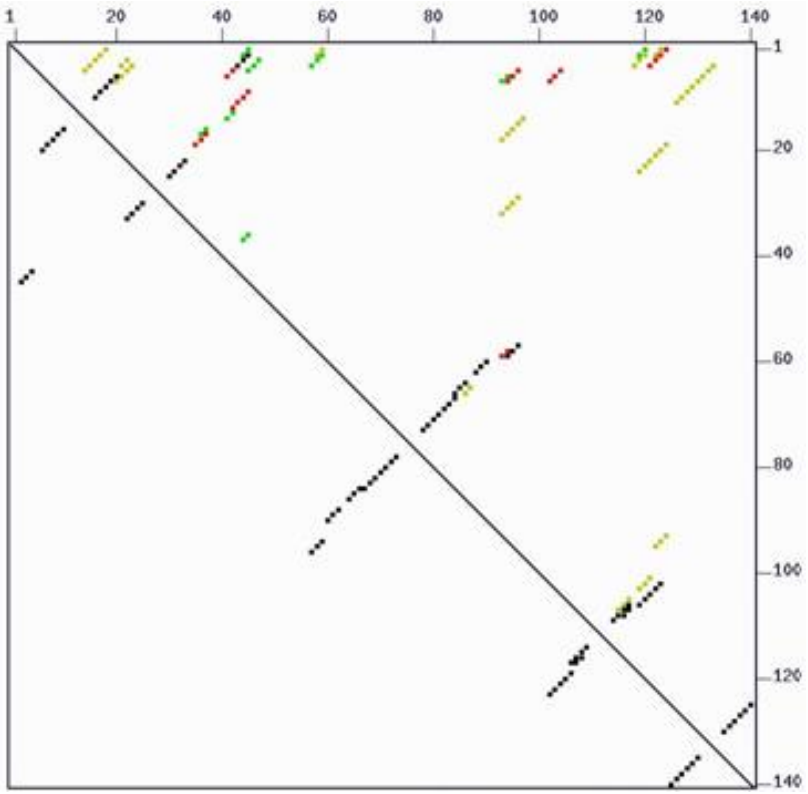
Consider whole sequence

- there may be many structures which are almost as good (slightly sub-optimal)

Treat in terms of probabilities

Probabilities





- lower left – best structure
- upper right – probabilities of base-pairs

State-of-the-art predictors

Related sequences from other species fold the same way

Procedure

- collect closely related RNA sequences from data bank
- try to fold all simultaneously

Kinetics..

Imagine you can predict 2D structures

- are you happy ?

Two possible scenarios

- kinetic trapping
- slow formation

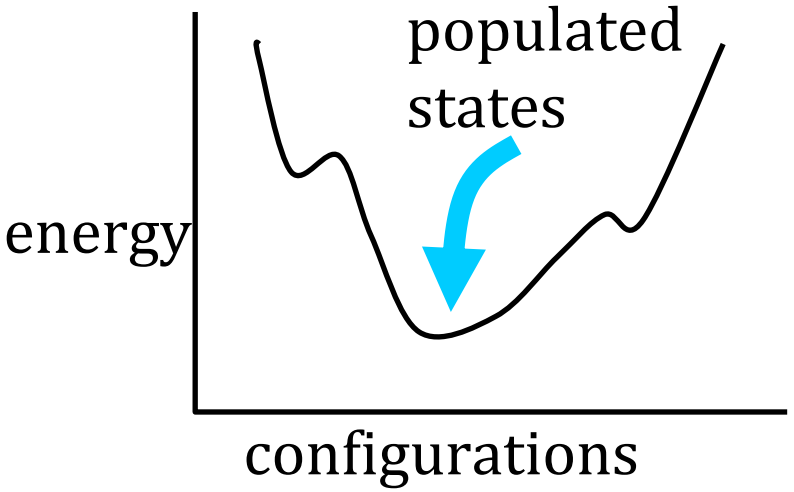
Kinetic trapping

Term from protein world

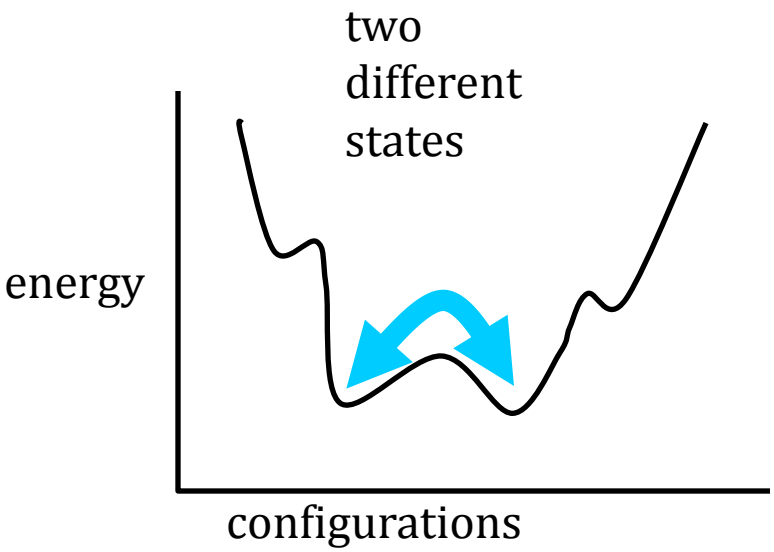
Wherever the molecule is

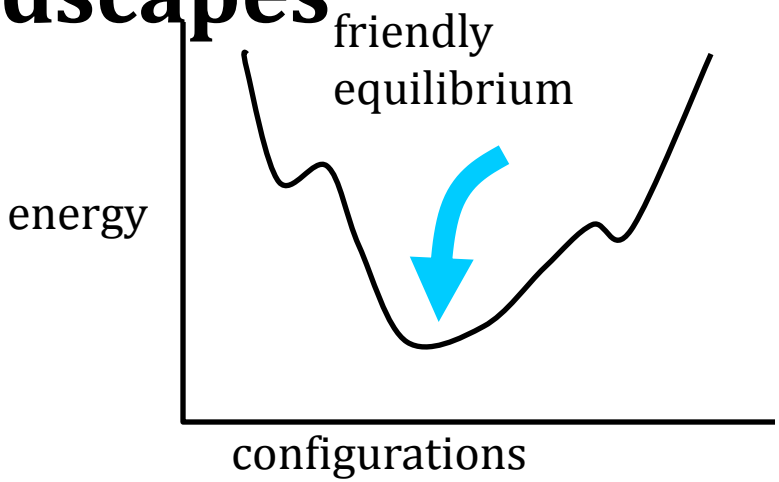
- it will probably go to energetic minimum

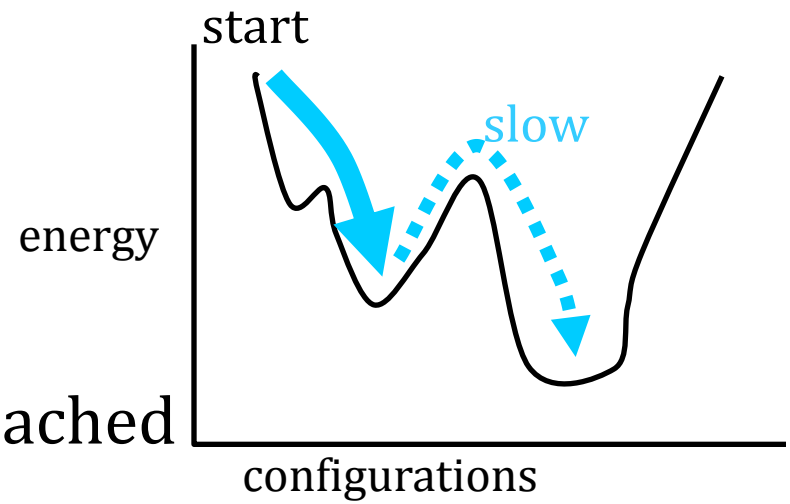
- less friendly landscape



Energy landscapes







If barrier is too high, best conformation may never be reached

How real is the problem ?

Consider base of type G

- there are many C's he could pair with
- only one is correct

- there are lots of false (local) minima on the energy landscape

Landscapes / kinetics

Can one predict these problems ?

- not with methods so far

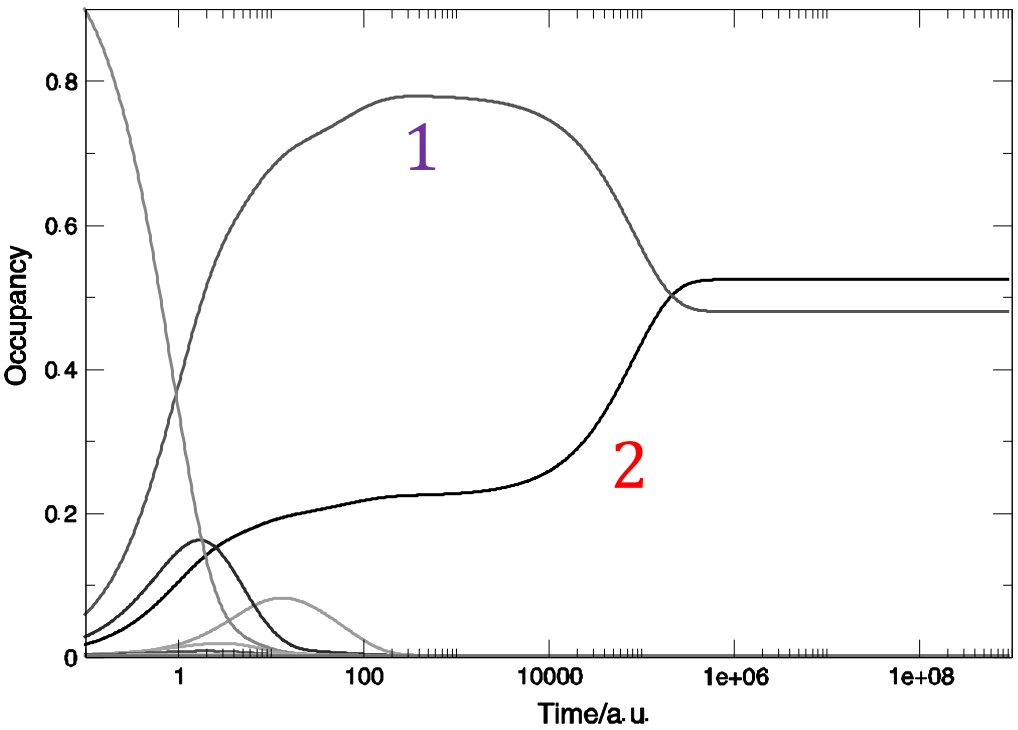
Try with simulation methods

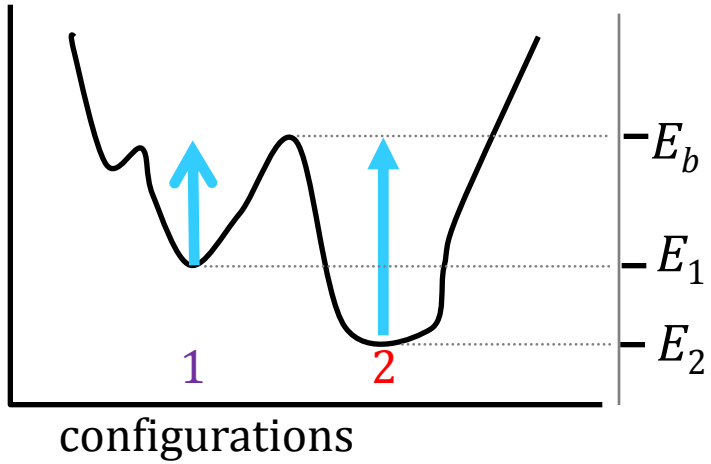
  - Monte Carlo / time-based methods

- start with unfolded molecule
- use classic methods to get a set of low energy predictions
- simulate folding steps
  - measure amount of each good conformation with time..

Example calculation

- conformation 1 forms rapidly
- conformation 2 slowly forms
  - conformation 1 disappears



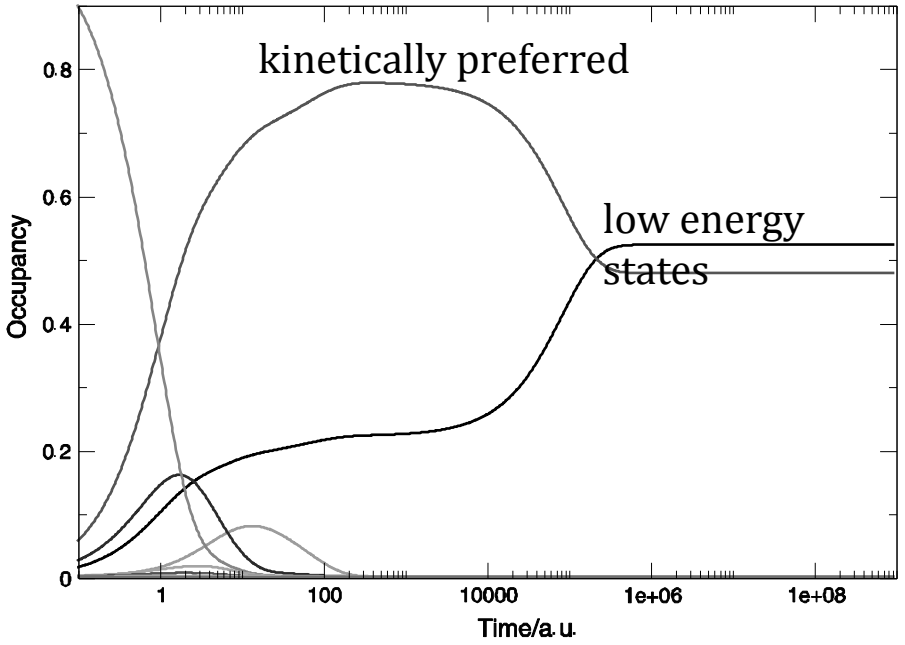


Implications

What if RNA is degraded ?

Molecule disappears before it finds best conformation

"kinetically preferred" conformations may be more relevant than best energy



summary

Tertiary structure very important (binding of ligands)

2D (secondary structure calculations)

- fast
- limits structures one can predict (no pseudoknots)
- predictions are not reliable
- used everywhere in literature (coming seminars)

You may lose anyway (kinetics)