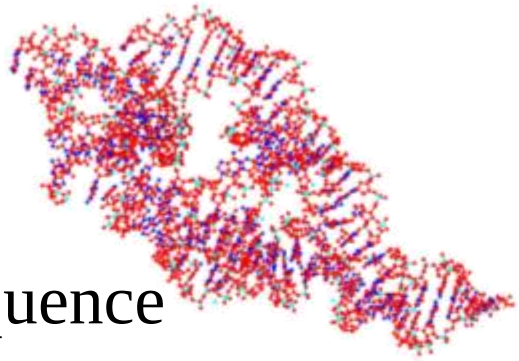
RNA structure, predictions

Themes

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

- This handout for today and next week

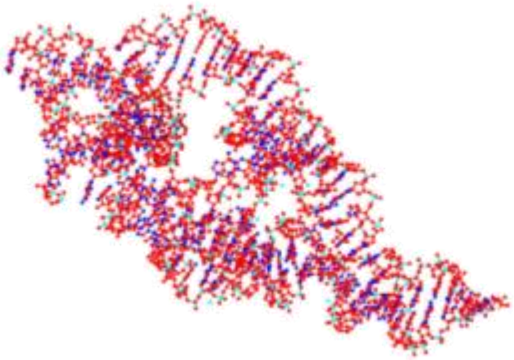
Structure



Analogy to proteins

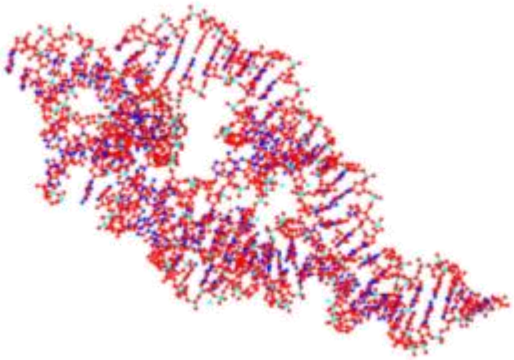
- Proteins
  - common belief – unique structure for sequence
  - 20 amino acids, many specific interactions
    - hydrophobic, charged, big, small, …
    - hydrophobic core
  - 8×10^5 structures in databank
- RNA
  - $< 10^3$ structures in databank
  - 4 bases
    - 2 bigger, 2 small
  - less specificity ? fewer unique structures

Protein vs RNA



- middle of proteins
  - hydrophobic core
    - soup of insoluble side chains
- middle of RNA
  - specific (Watson-Crick) base pairings
  - other base pairs
  - much more soluble…

RNA – how important is 3D structure ?

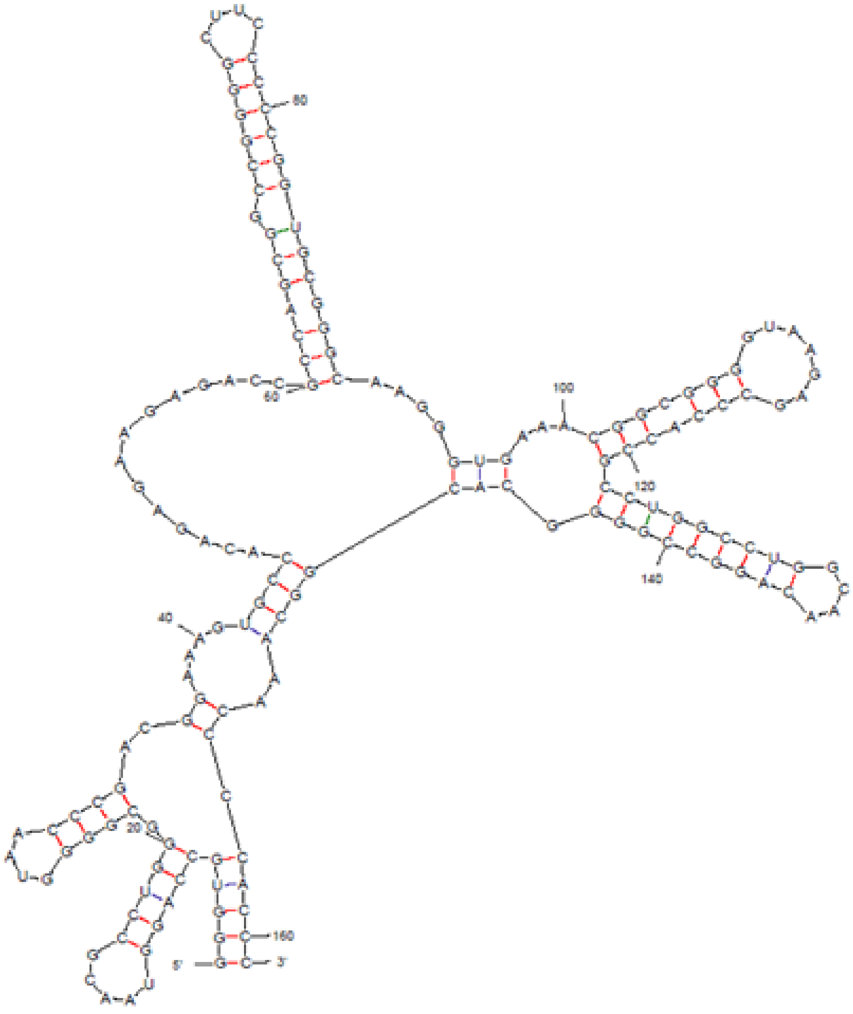


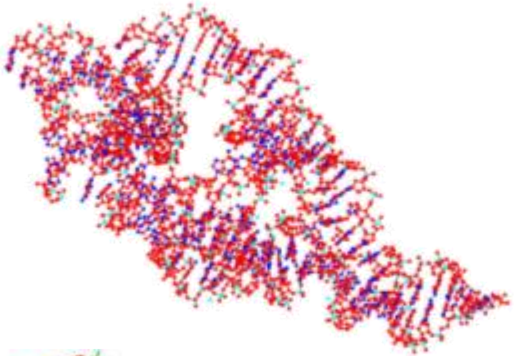
- primer design, blocking DNA, ..
  - only think of base pairs

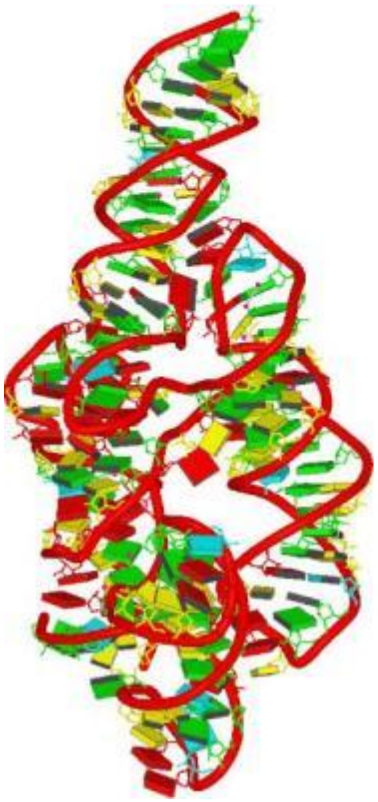
- binding of ligands (riboswitches. ribozymes)
  - totally dependent on 3D shape – where in space are functional groups

How realistic is 2D ?

- 3D versus 2D (1u9s)



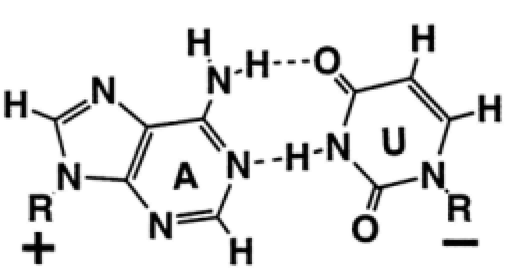


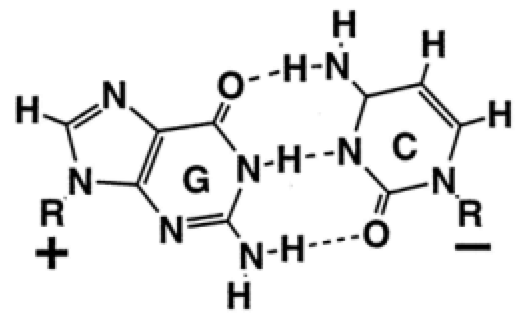


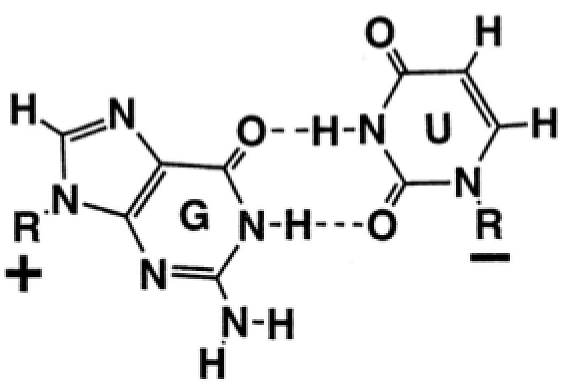
2D why of interest ?

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)

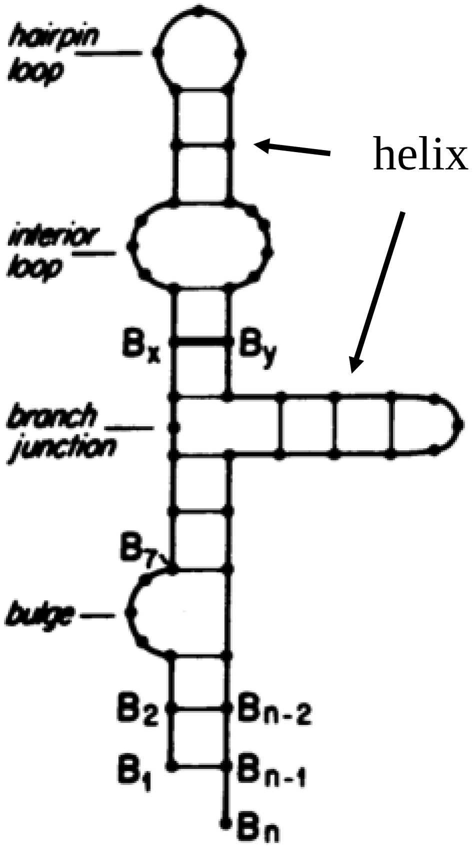
2D why of interest ?

3. Claim - RNA folds hierarchically

nearby bases fold first, later overall structure

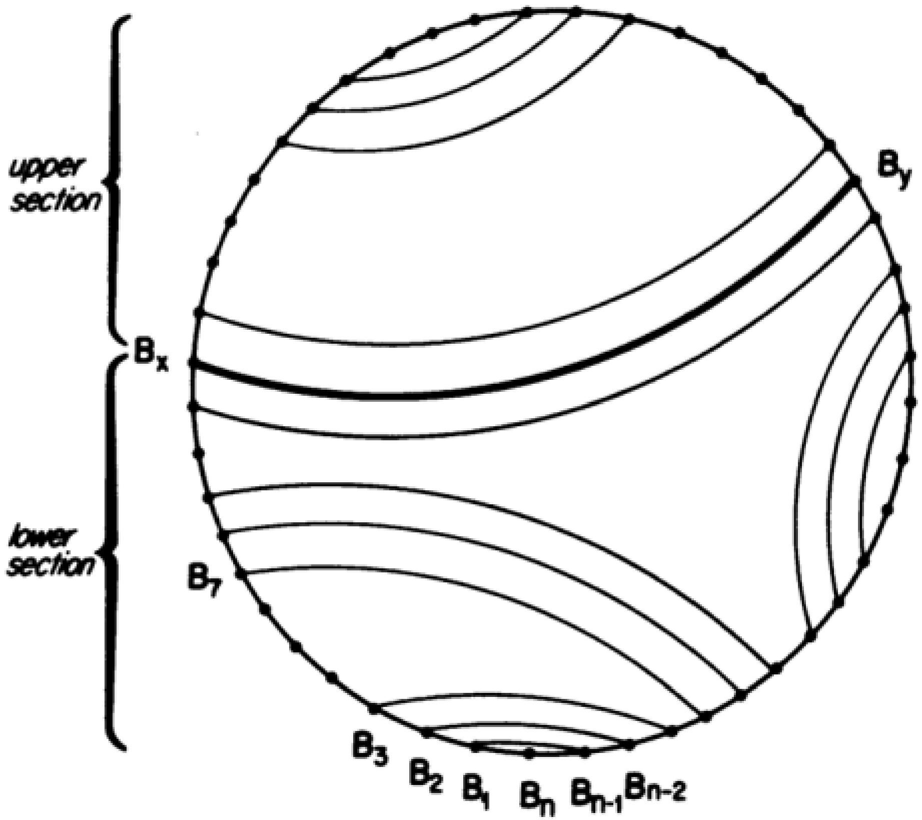
- evidence not clear
- much contrary evidence in protein world
- plausible in RNA world ?
  - RNA double strand helices are believed to be stable
  - contrast with proteins – isolated α-helices and β-strands are not stable in solution
- useful ?
  - if true, then 2D (H-bond pattern) prediction is really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

Four representations of flat RNA



1. conventional

- + on next slide

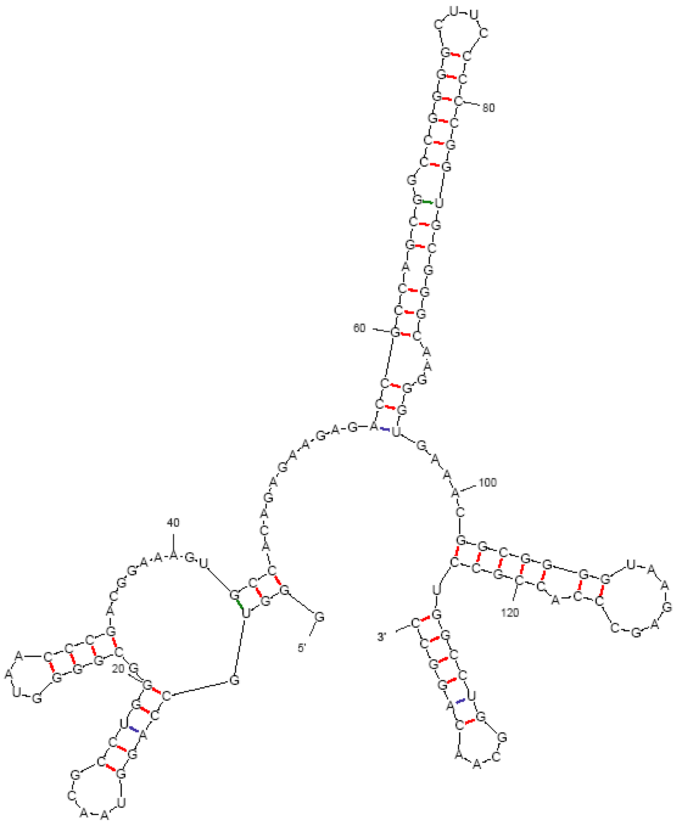


2. Nussinov's

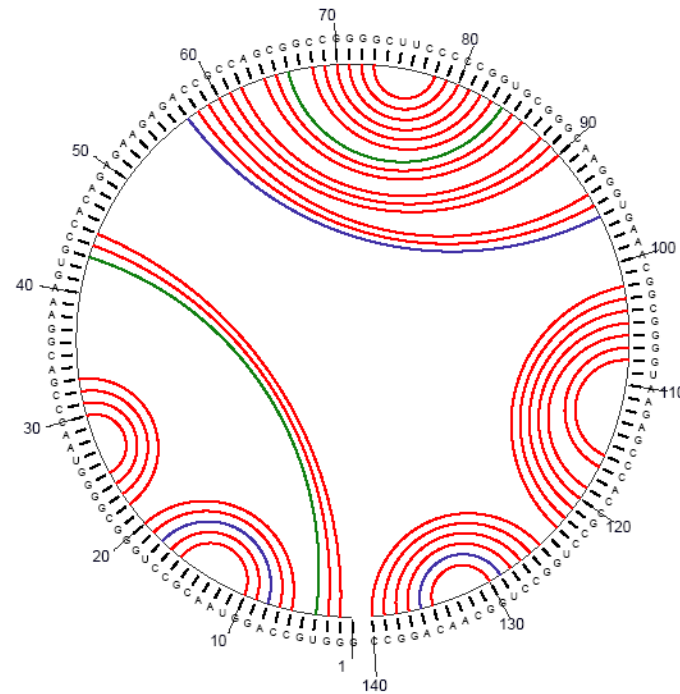
- write down bases on circle
- arcs (lines) may not cross

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

Four representations of flat RNA



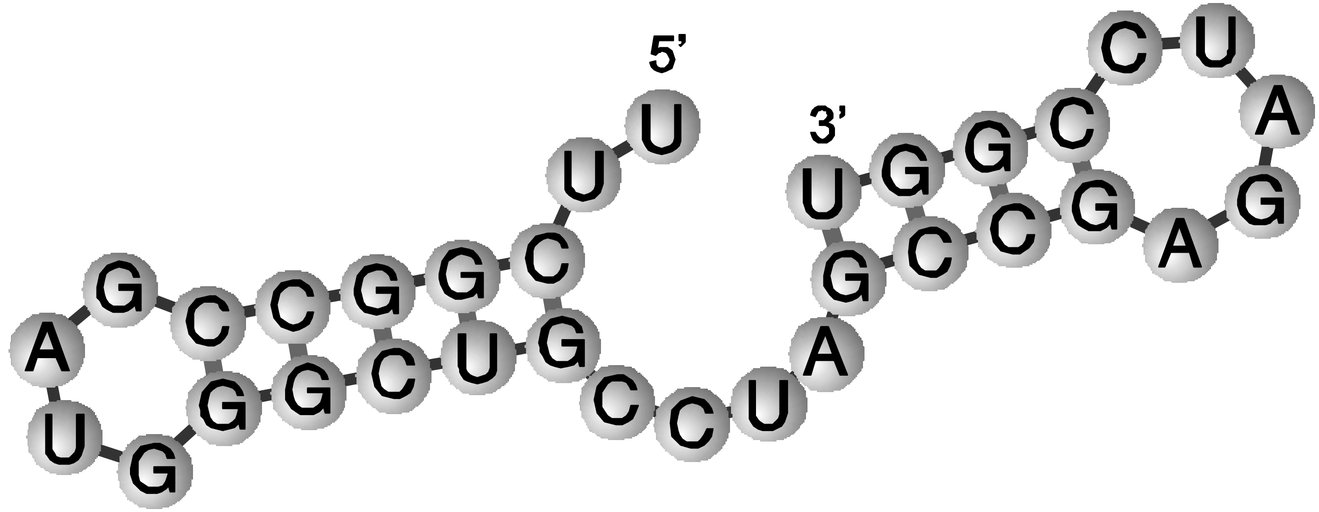
1. conventional representation



2. Nussinov's circle

- same features in both plots

Parentheses



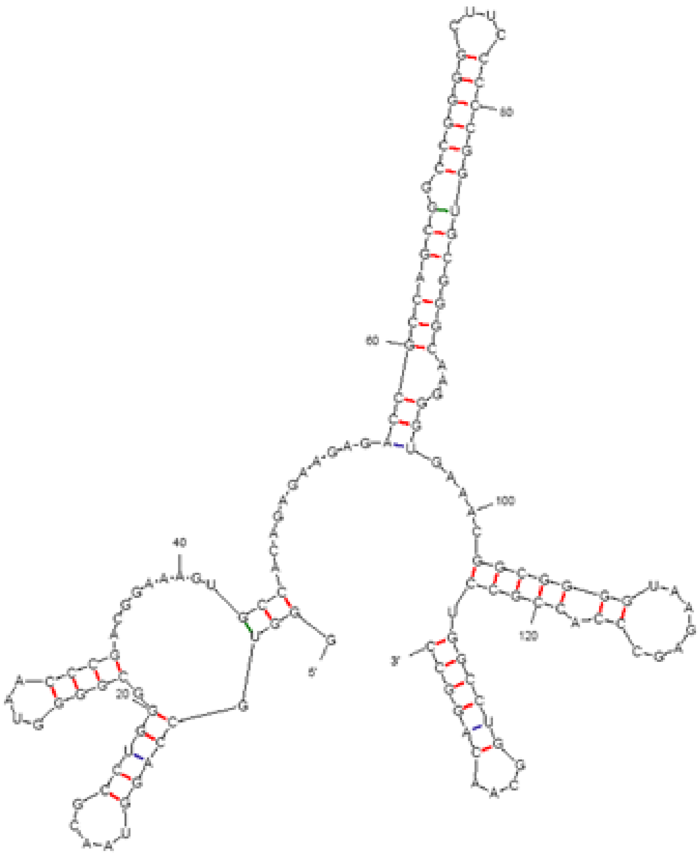
- 3. parentheses – most concise

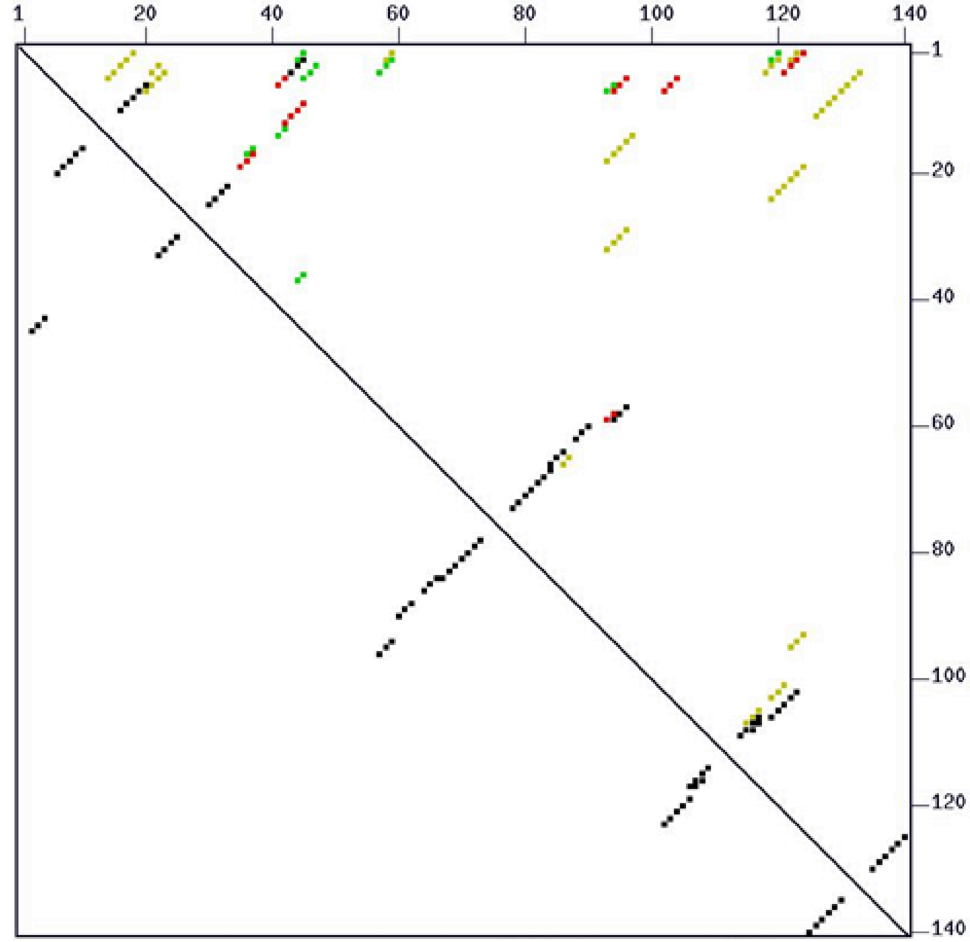
  `..(((((....)))))....((((.....))))`

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

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

Dot plots



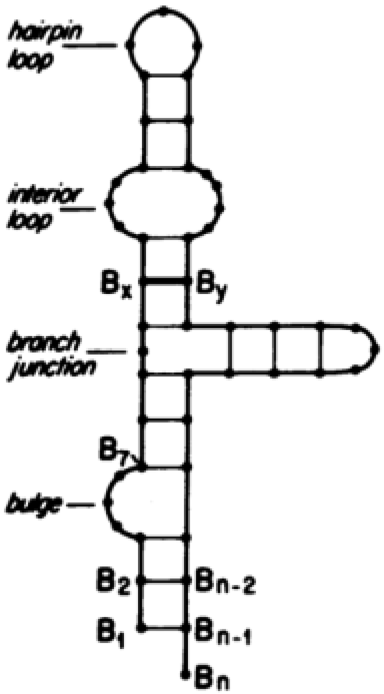


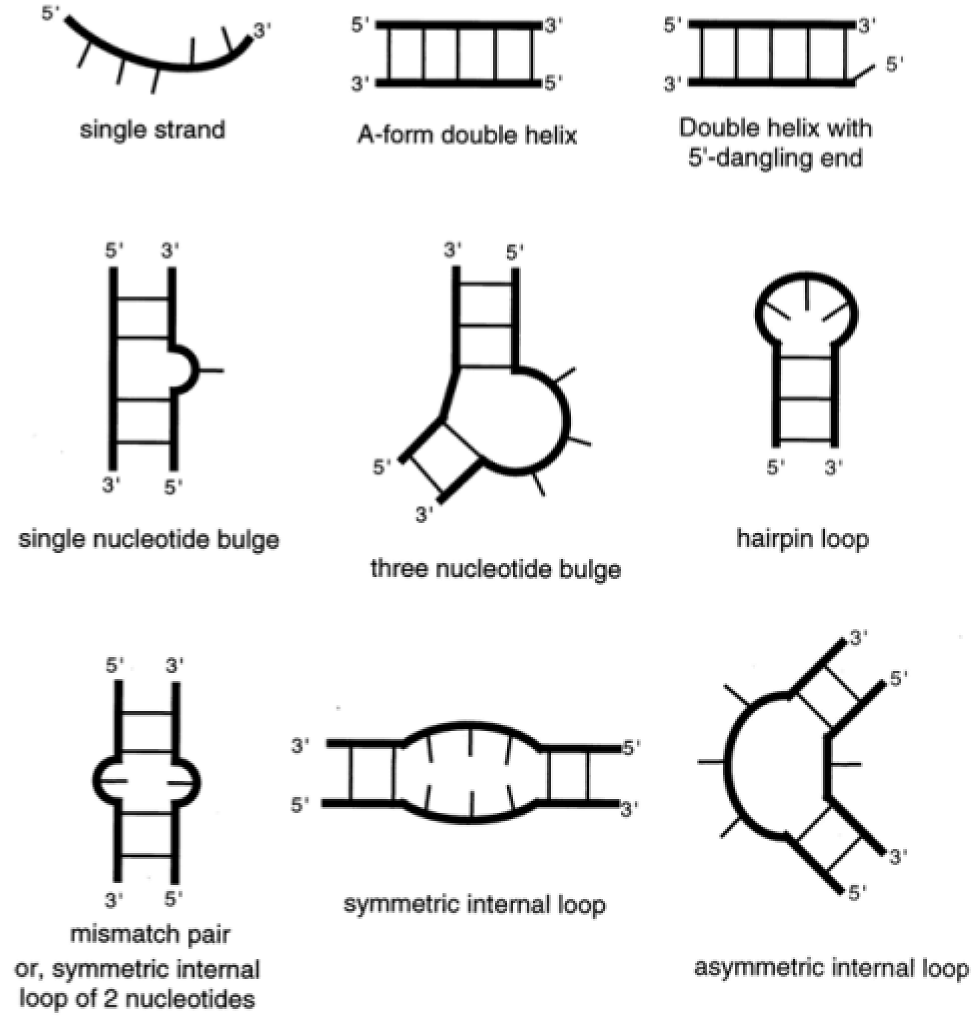
4. Dot plots

- 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



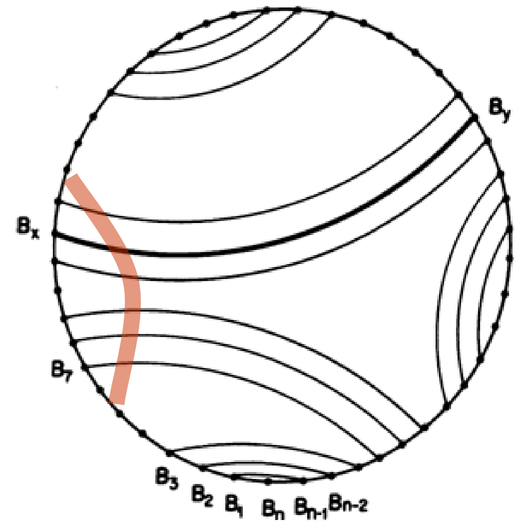


- for explanations later
  - branch junction
  - hairpin loop
  - bulge
  - interior loop / mismatch

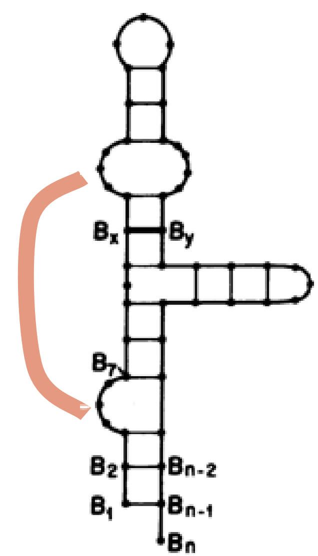
from Nussinov, R., Jacobson, A.B. Proc. Nati. Acad. Sci. USA, 77, 6309-6313(1980)
from Burkard, M.E., Turner, D.H., Tinoco Jr., I., in The RNA World, 2nd Edn, eds Gesteland, RF, Atkins, JF Cold Spring Harbor Laboratory Press (1999)

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 ?

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

Predicting secondary structure

- how many structures are possible for n bases ?

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

for some constants c and $d \approx 1.8$

- exponential growth

- problem can be solved
  - restriction on allowed structures
  - clever order of possibilities

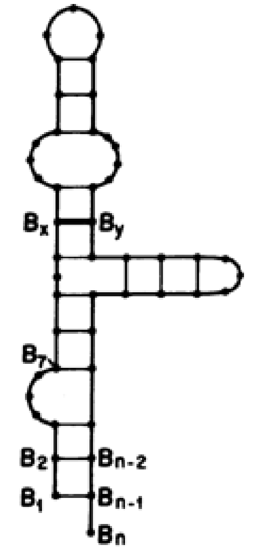
Best 2D structure (secondary)

- scoring scheme :
  - each base pair scores 1 (more complicated later)

- Problem
  - some set of base pairs exists – maximises score
  - crossing base pairs not allowed
- our approach
  - what happens if we consider all hairpins ?
  - what happens if we allow hairpins to split in two pieces ?

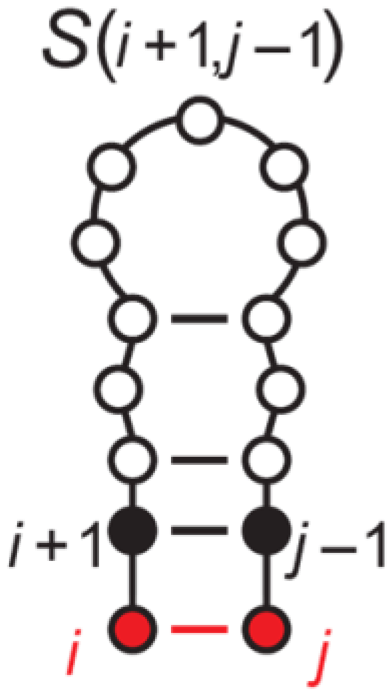
Philosophy

- 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

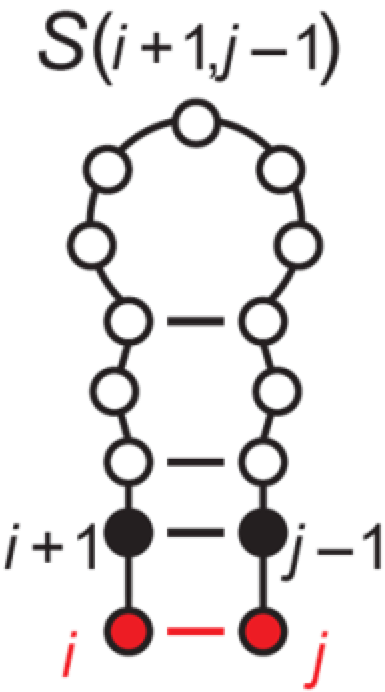
- start by looking for best possible hairpin
- idea
  - if we know the structure of the inner loop
    - we can work out the next
  - if we know the black parts
    - we can decide what to do with the red *i* and *j*



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* ?

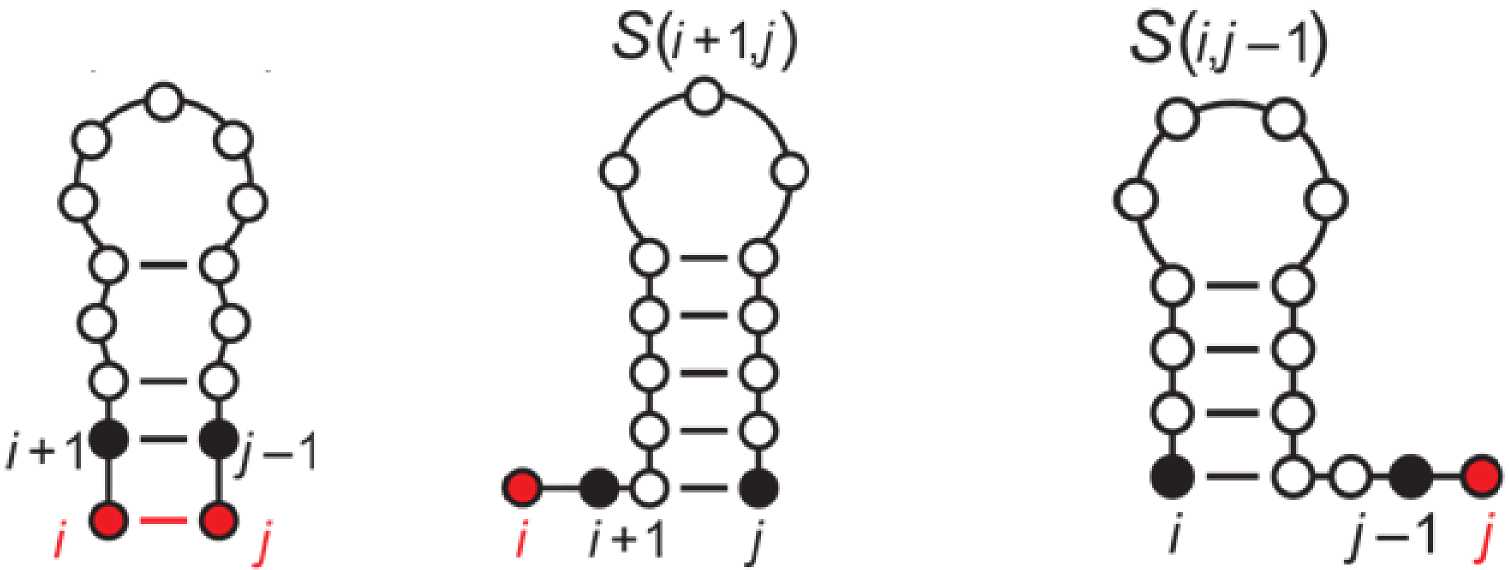


- maybe *i* should pair with *j*
- maybe there is a better *j* later

- what possibilities must one consider ?

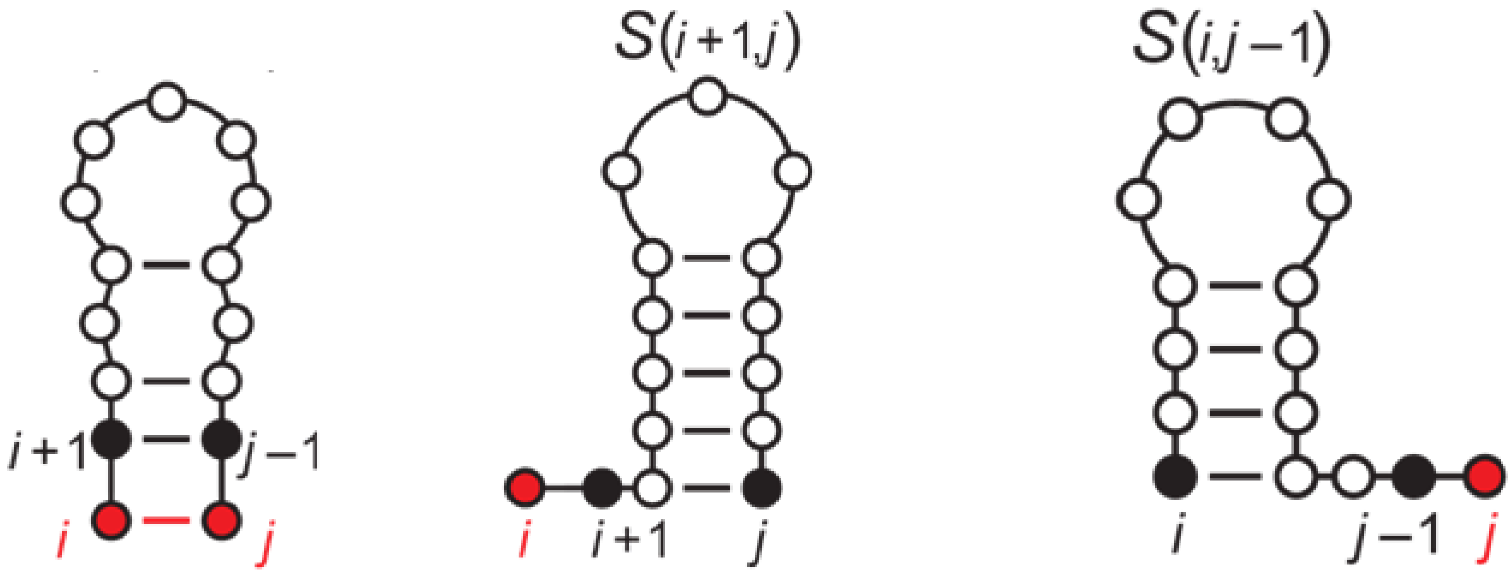
Optimal hairpins

- 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

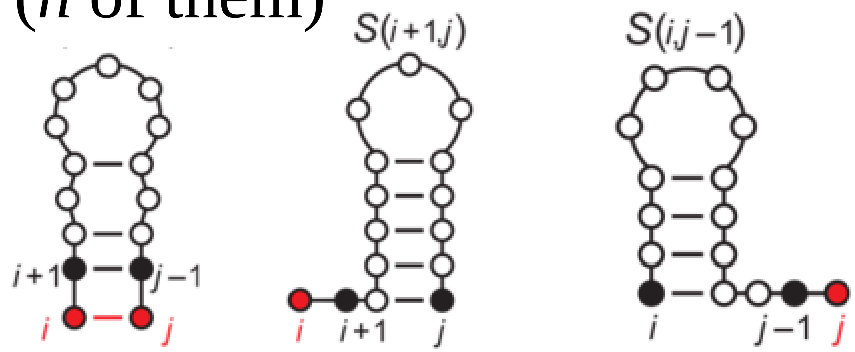


- order of steps
    - start by finding best local loops/pairs
    - move outwards

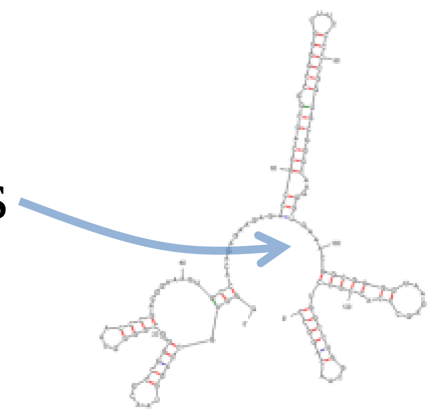
- consequence
    - base pairs will never cross - important

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

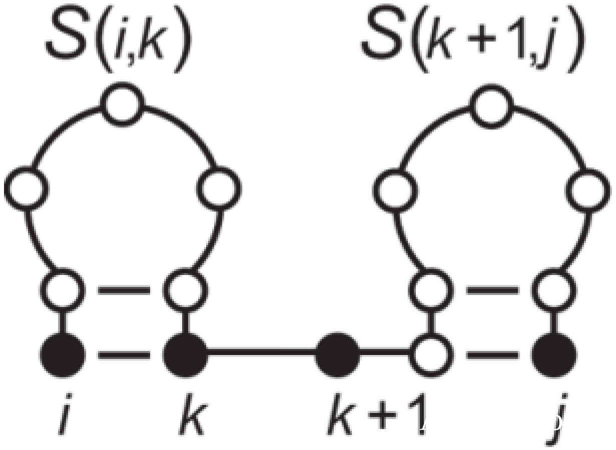


- What have we done ?
  - best organisation of hairpins
    - with best position of bulges and gaps
- Cannot yet split a chain into multiple hairpins



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 expensive ?



cost of predicting structure..

- for each *i*
  - test each *j*
    - try each *k*
- $n \times n \times n = O(n^3)$
- not really so simple
  - very fancy order of steps (dynamic programming method)
- very severe limitation (pseudoknots later)

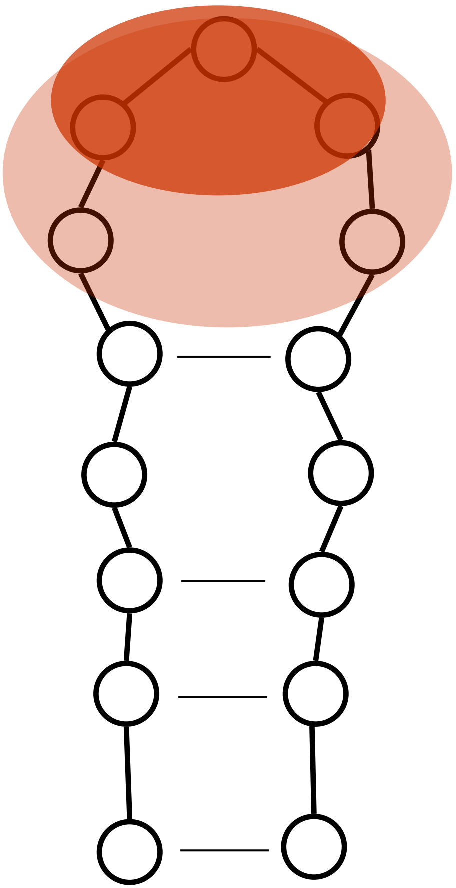
- In principle…
  - for a given sequence, can find the best arrangement bases
- needs more sophistication

Scoring schemes

- 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
- change the scoring scheme – improvement..
  - count H-bonds
- still not enough

non-base pair complications

- First approximation
  - each H-bond is independent of neighbours
    - all GC (or AU or GU) pairs are the same
- Other factors
  - loops and stacking..



- Consider unpaired bases
  - counted for zero before
  - compare loop of 3 / 5 / ..
- do these bases
  - interact with each other ? solvent ?
  - energy is definitely $\neq 0$

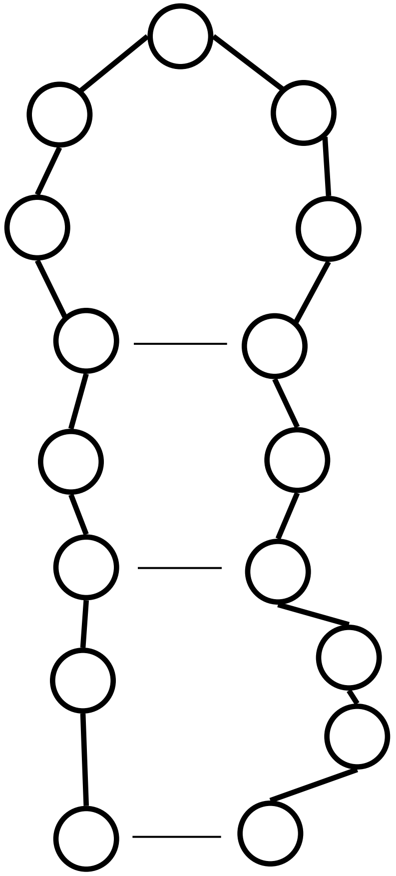
non-base pair complications

Unpaired bases

- one basepair bulge
  - distorts helix / costs energy at backbone
- two / three basepairs ?

How to treat

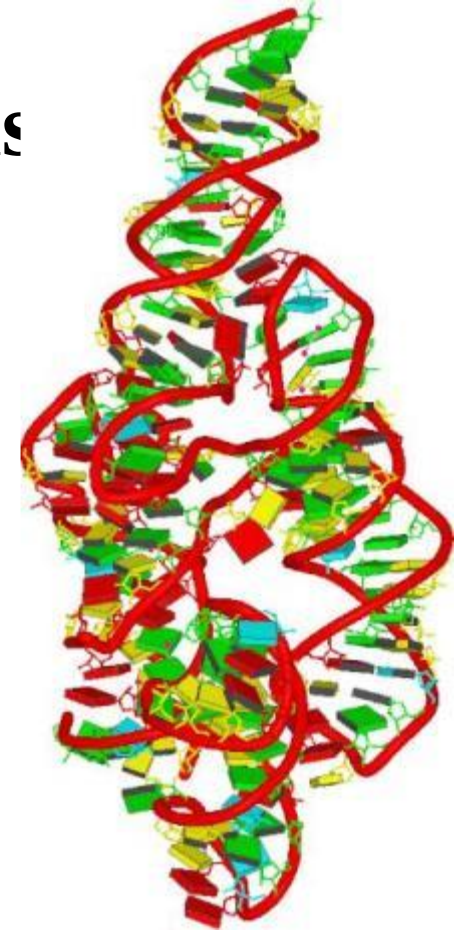
- like gap penalties in protein alignments
- when considering *i*, *j* pairs, add in penalties for bulges



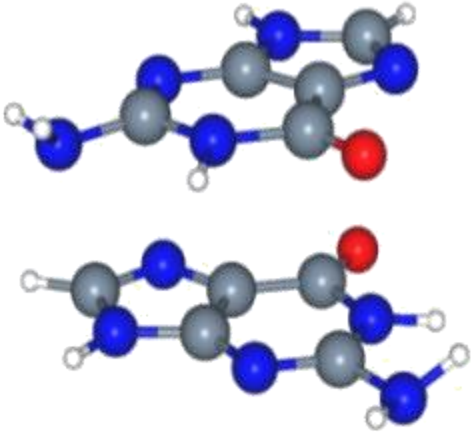
How much ?

- later

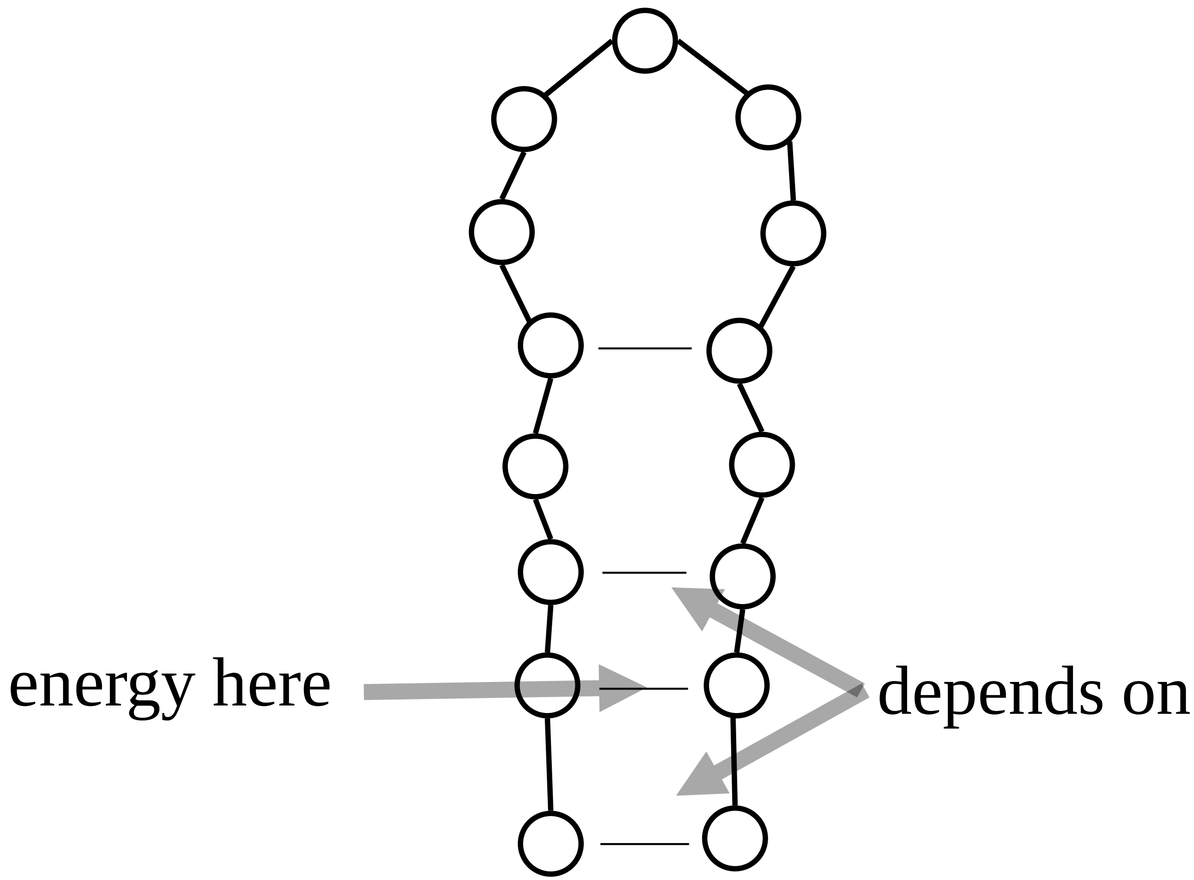
non-base pair complications



- Assumption: each basepair is independent
- $S(i,j)$ = base-pair + $S(i+1, j-1)$
- valid ?
  - consider all the interacting planes
    - partial charges, van der Waals surfaces



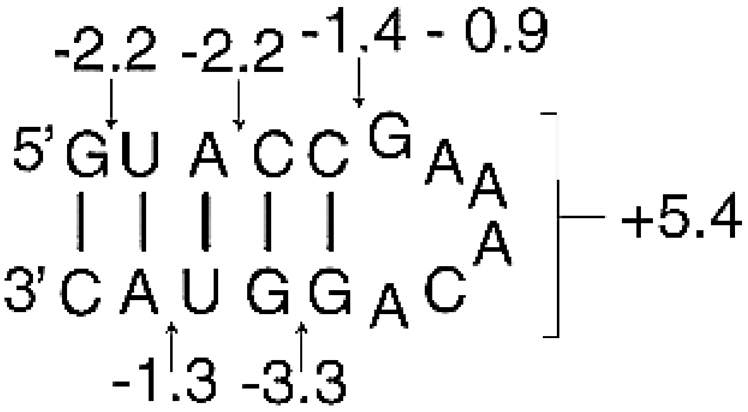
non-base pair complications



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

Nearest neighbour model

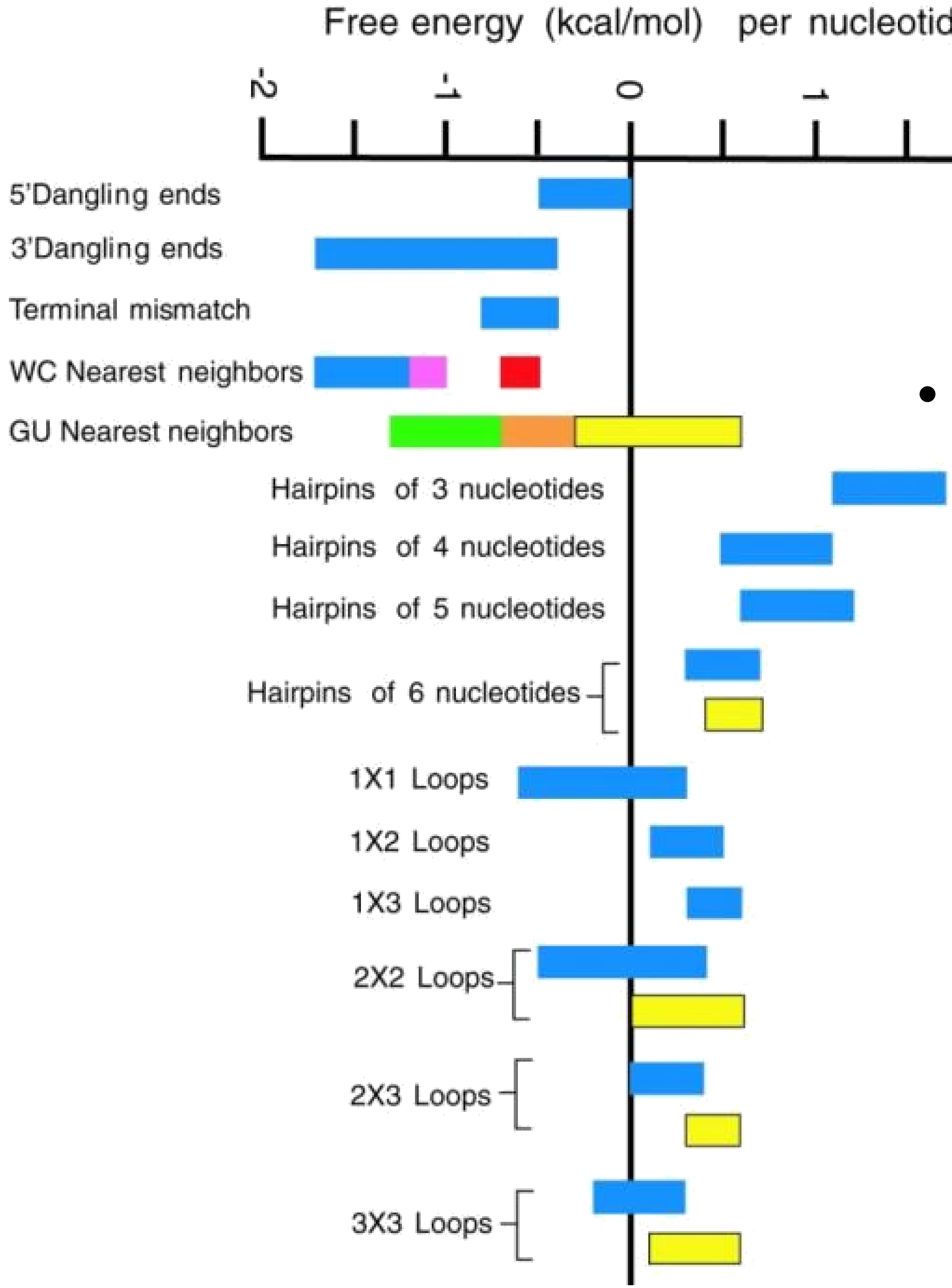
- Previously we added
  - GC + UA + AU + …
- Now
  - (GU/CA) + (UA/AU) +..



  - terminal loop costs 5.4 kcal mol^{-1}
  - where do numbers come from ?

Nearest neighbour model

- parameters..
  - model is not perfect – a (GU/CA) pair will depend on its environment
  - best guesses
    - make small helices, measure melting temperatures of related sequences
    - ACTGACTG vs ACTAACTG tells you about TG vs TA
    - make loops of different sizes and measure melting temperatures



- values
  - are not precise
  - depend on context

  - colours are for different kinds of neighbours

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

Score summary

simplest	count base pairs
medium	count H-bonds
complicated	nearest neighbour model pairs of pairs, loops, ends, …

- how accurate ?

Reliability

- how accurate ?
  - too many factors, sequence environment, possible tertiary effects
  - maybe 5 – 10 % errors
- how good are predictions ?
  - maybe 50 – 75 % of predicted base pairs are correct
- why so bad ?

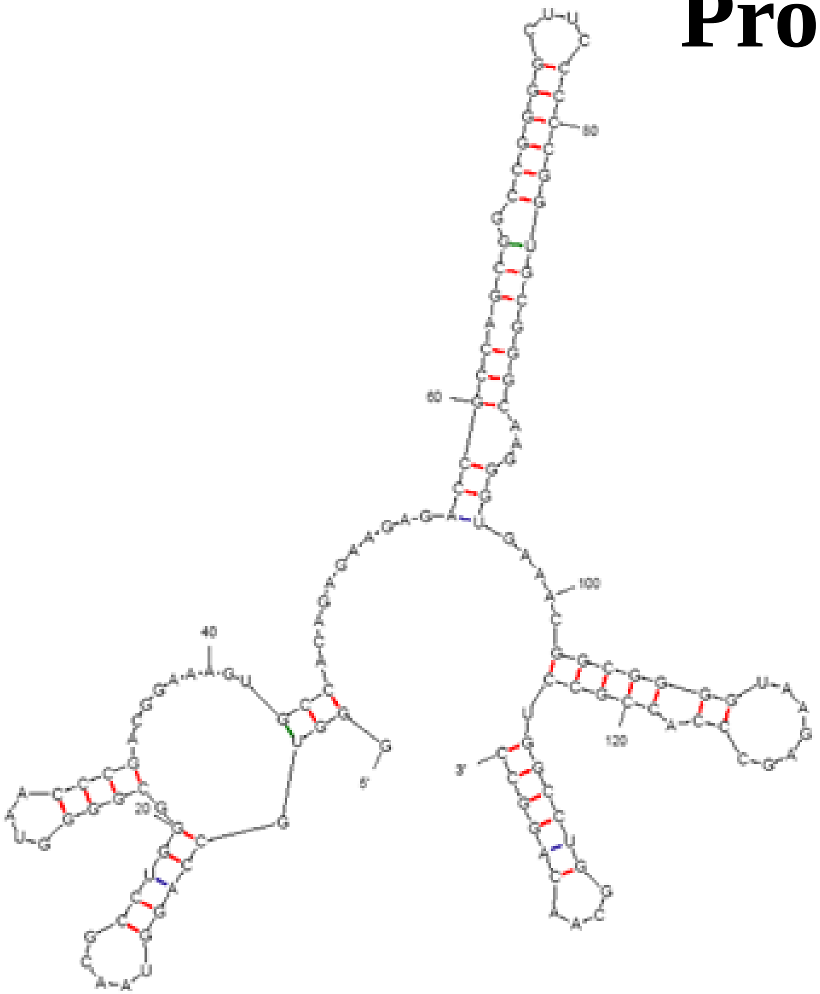
Reliability

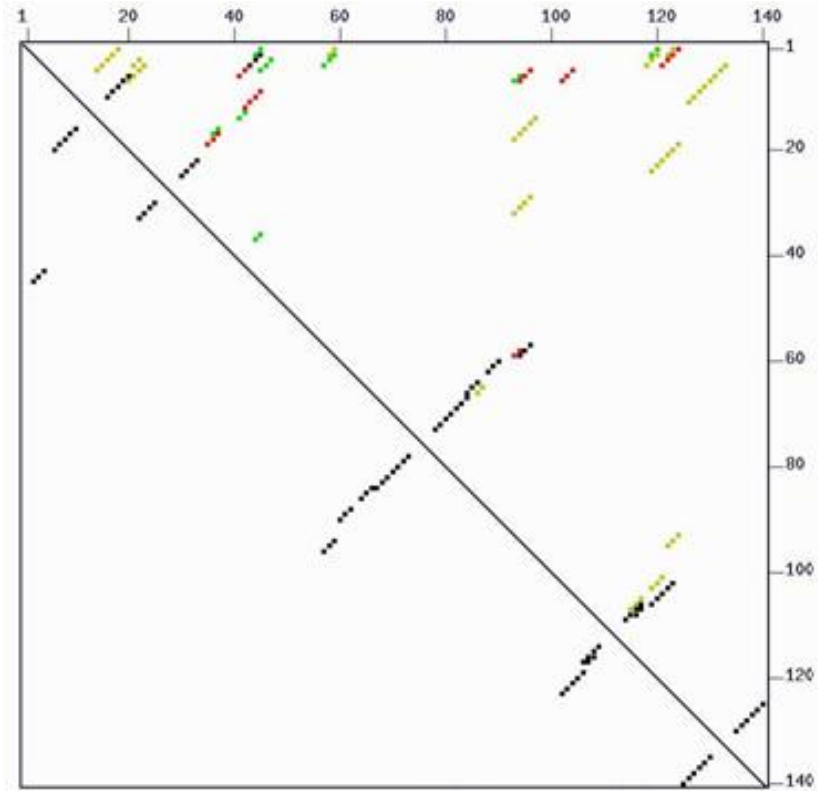
- Remember nature of RNA
  - only 4 base types
- think of an "A"
  - wants to pair with a U
  - there are many many U's
- think of any base
  - many possible good partners
- consider whole sequence
  - there may be many structures which are almost as good (slightly sub-optimal)
- importance of sub-optimal solutions…

Reliability

- for some sequence
  - there are 999 wrong answers with good energies
    - + 1 correct answer
  - add in error to all the values and pick the most negative
    - probably will not be the correct one
- can they be improved ?
  - work with sets of aligned sequences
- consequence..
  - much effort in finding non-optimal answers
  - remember probability plots from earlier ?

Probabilities





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

Probabilities

- Have you met the Boltzmann relation ?
- probability p_i of being in state i

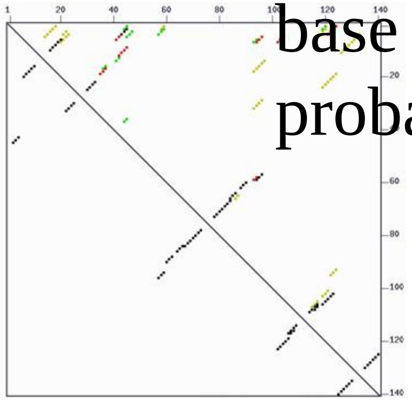
$$p_i \propto e^{-E_i/kT}$$

T	temperature
E	energy
k	Boltzmann constant

- i here is some base pair
- how is it calculated ? (not for exam)

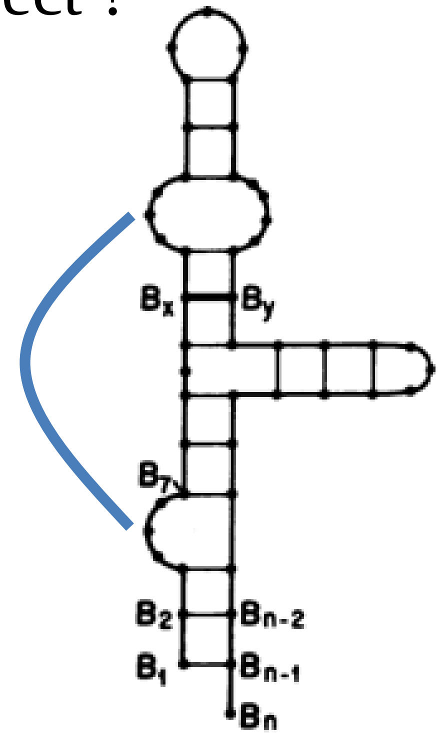
base pair probabilities

best base pairing



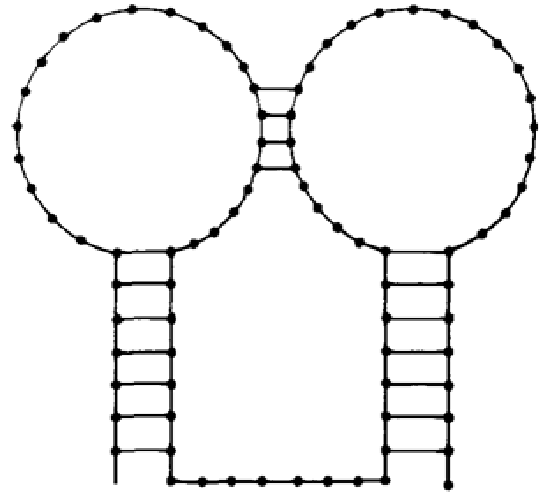
Problems

- 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
        - (definition of recursions)
      - structure of loops
      - assumption that energy is the sum of **enclosed** pairs



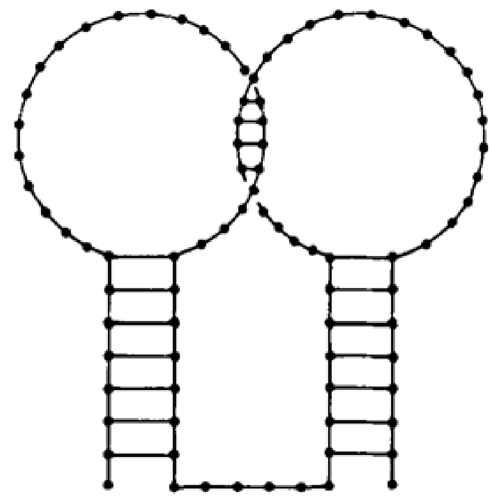
- General name … pseudoknots
  - why ?

Pseudoknots



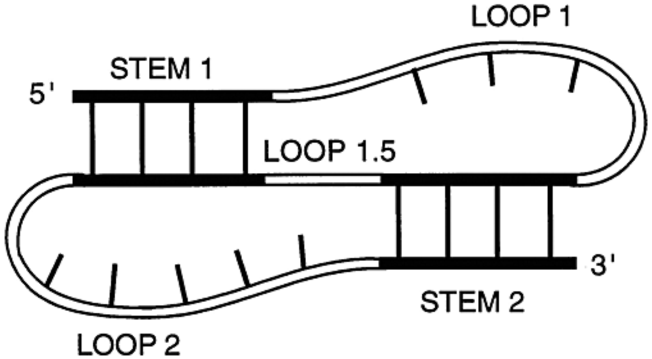
- pseudo-knot – not a knot
  - why the name ?

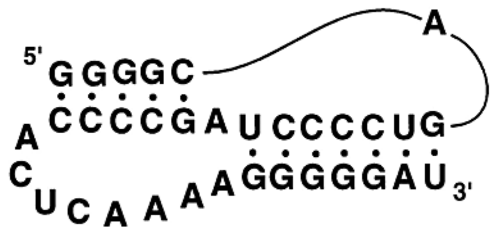
- topologically like a knot

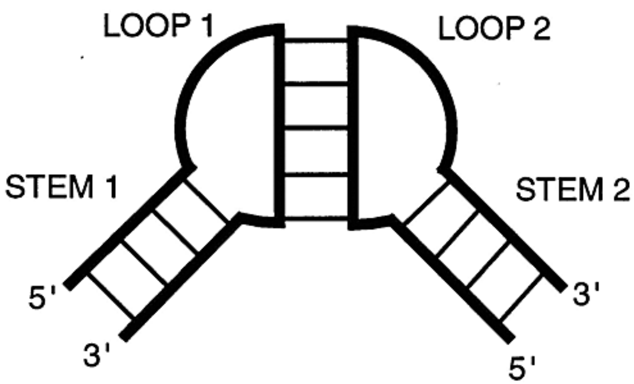


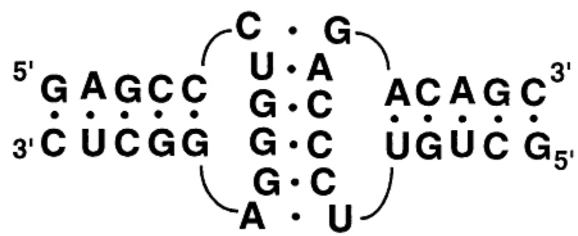
picture from Zuker & Sankoff, Bull. Math. Biol. 4, 591-621 (1984),
RNA secondary structures and their prediction

pseudoknots

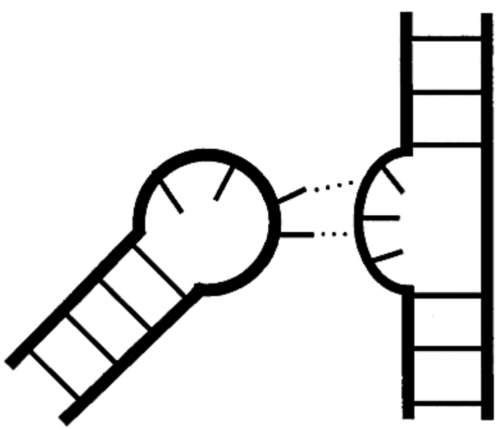


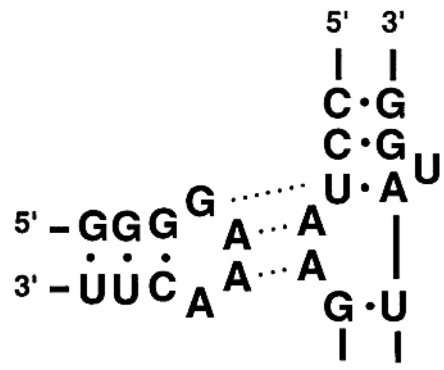






kissing hairpins





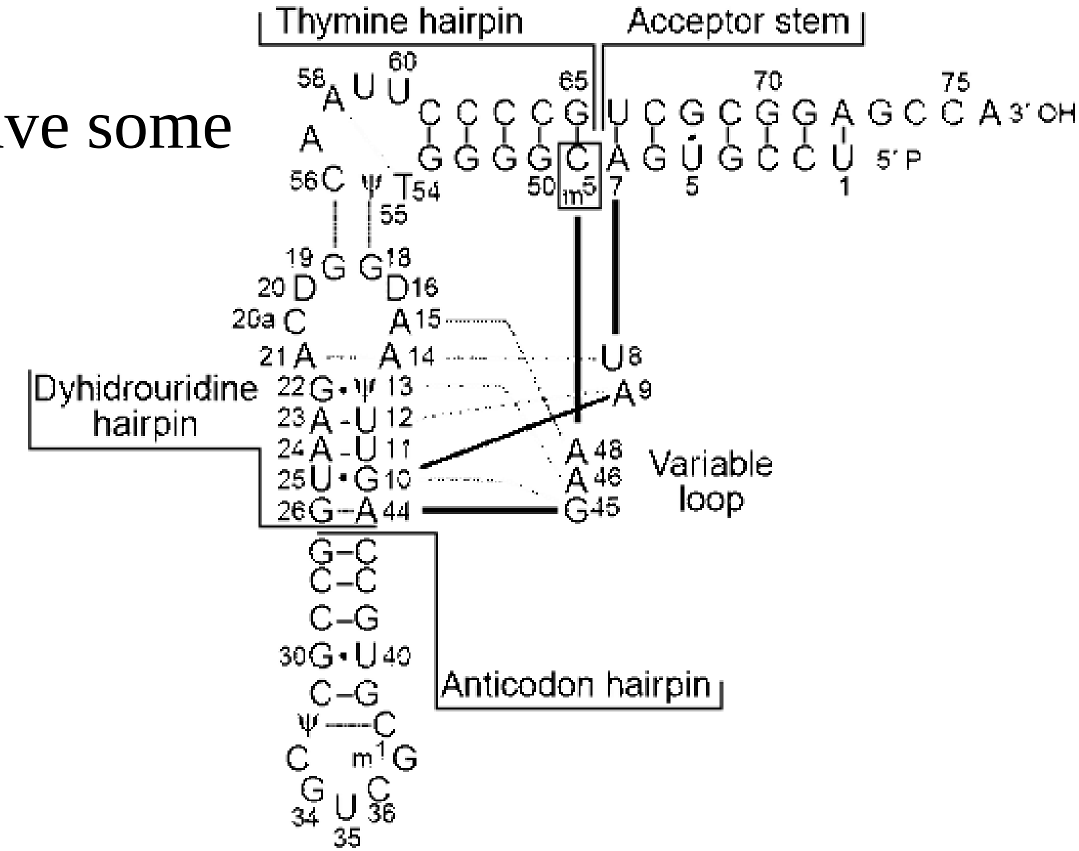
hairpin loop - bulge

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)

pseudoknots

Frequency of pseudoknots ?

- a few % of all H-bonds
- 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

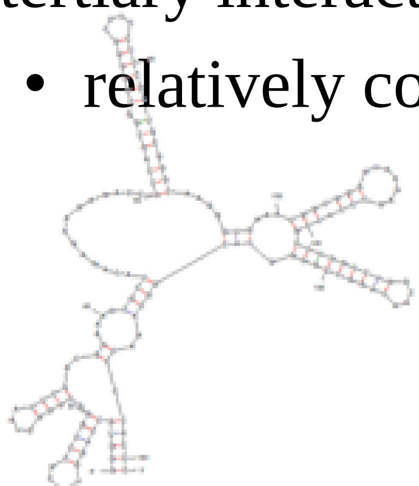
pseudoknot summary

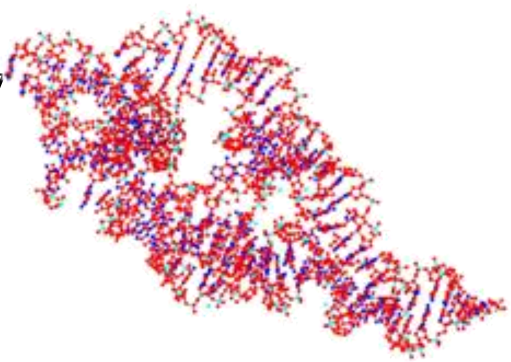
- fast algorithms cannot find pseudoknots
  - in order to go fast, the algorithms work in a special order
  - some base pairs come in "wrong" order
- more general problem
  - we have ignored tertiary interactions..

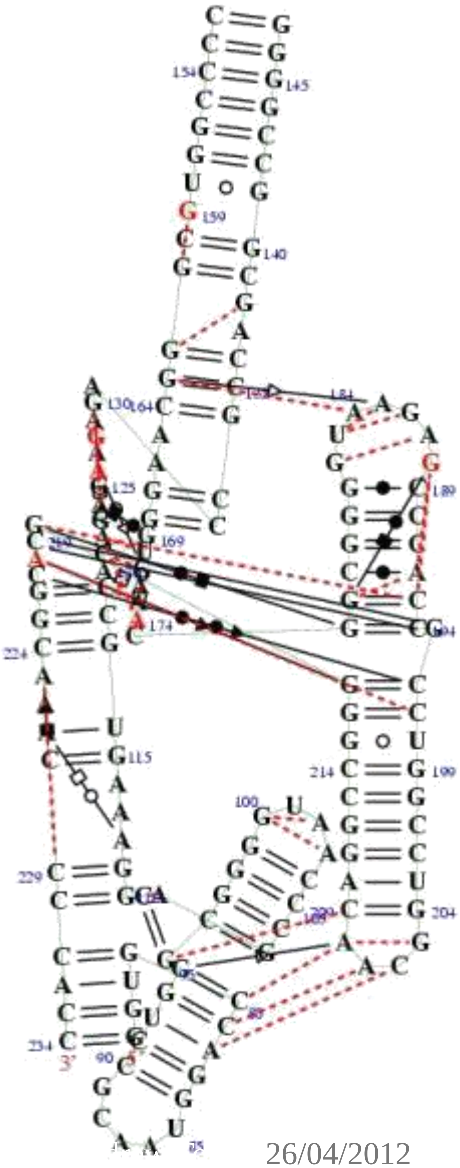
Tertiary interactions

- pseudoknots usually refer to classic H-bonding
- tertiary interactions could come in other forms
  - bases stacking
  - miscellaneous H-bonds
  - non-specific van der Waals

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





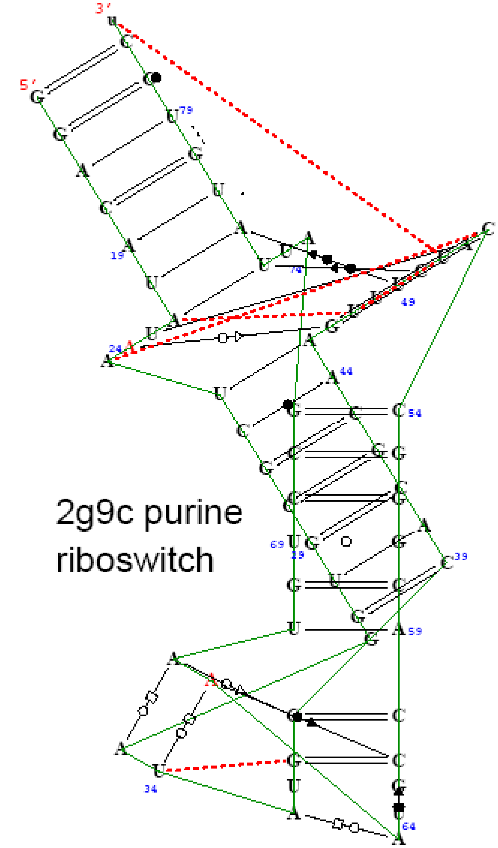


tertiary interactions from crystal, flattened

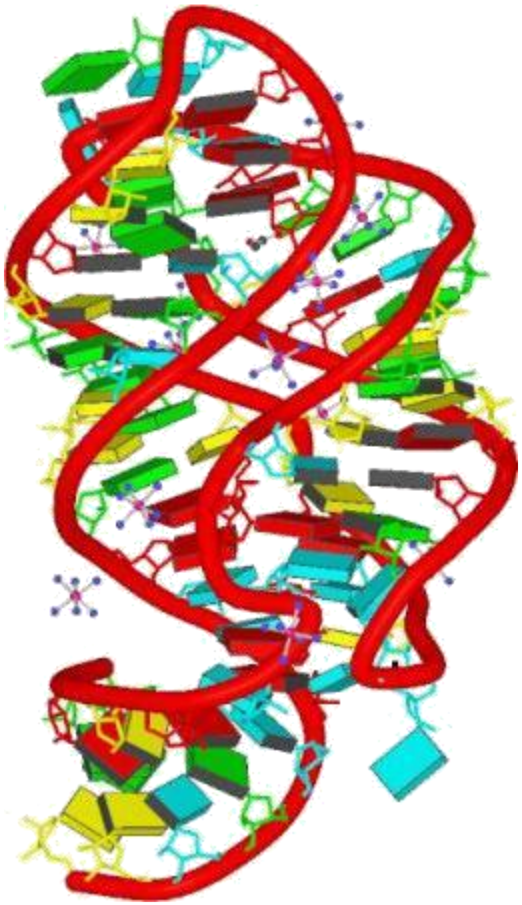
Pessimist view – all useless

- realistic, but nasty problems
- application – can we look for riboswitches ?
  - sequence where there is two different but good solutions
- realistic pictures

Horror 1



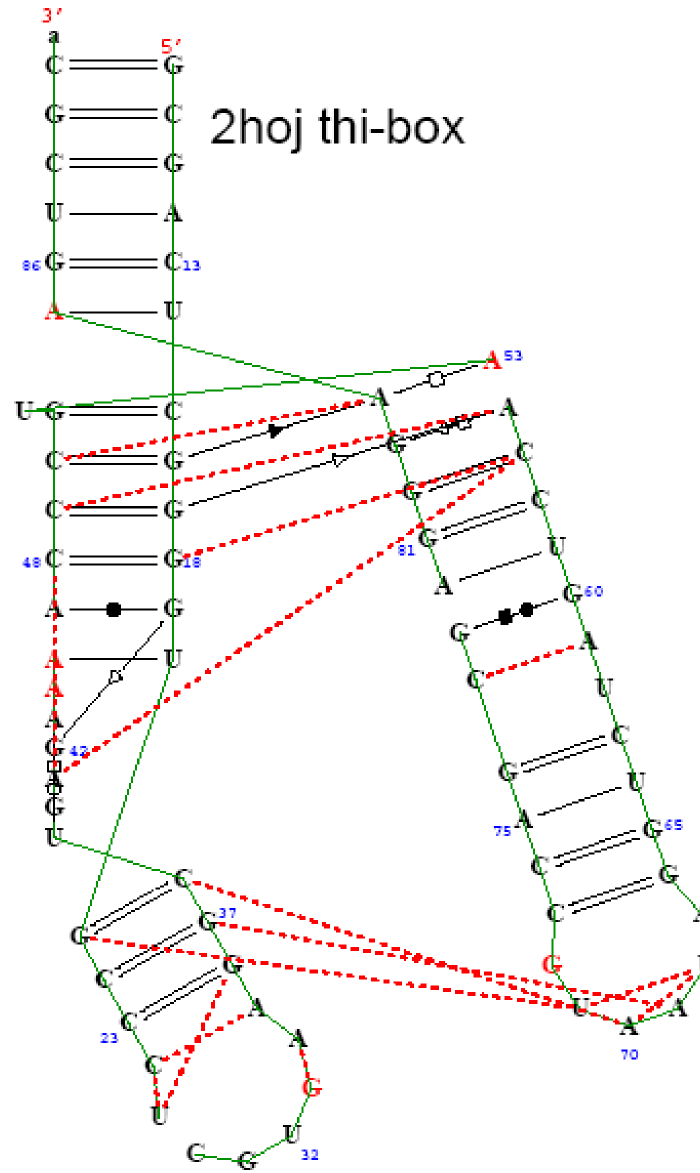
- 2g9c early riboswitch
  - 3D view – flat ?



- one conformation crystallised
- could you predict the other ?
- could you predict this structure ?
  - look at the number of strong interactions – not simple pairs

Horror 2

- same problem as before



3D predictions

- not practical
  - molecular dynamics simulations ?
    - not a friendly system – highly charged
    - too many atoms
    - interactions with metal ions
  - some claims of success

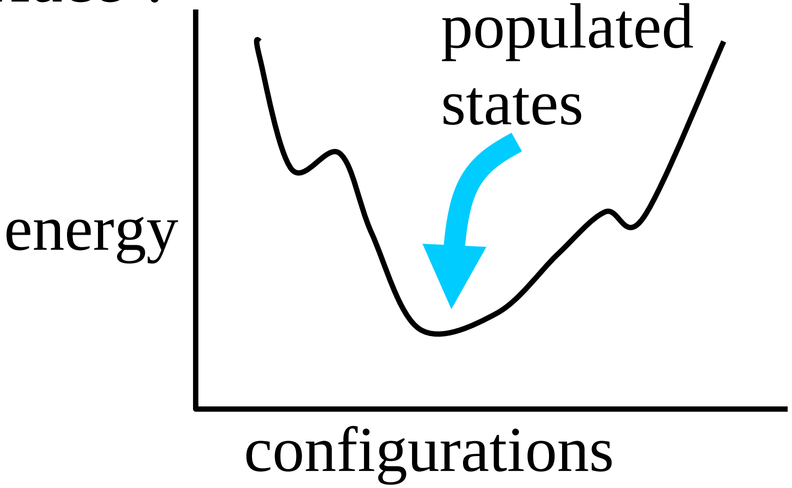
Kinetics..

- Imagine you can predict 2D structures
- do you win ?

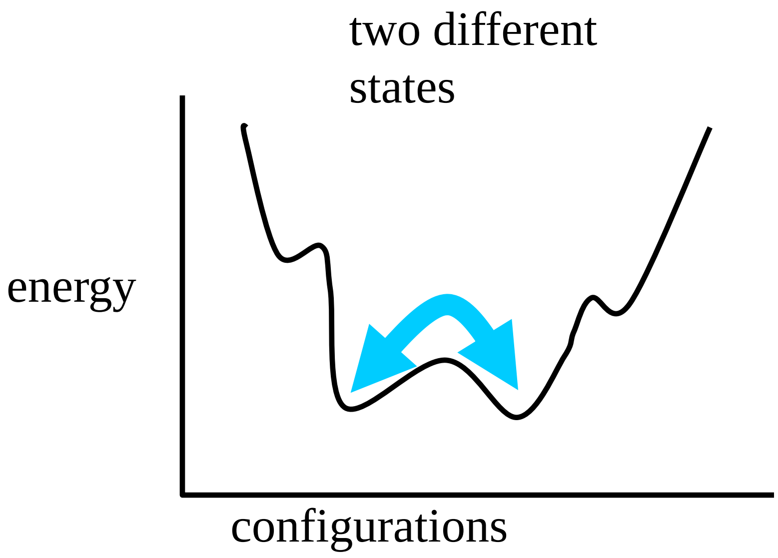
- two possible scenarios
  - kinetic trapping
  - slow formation

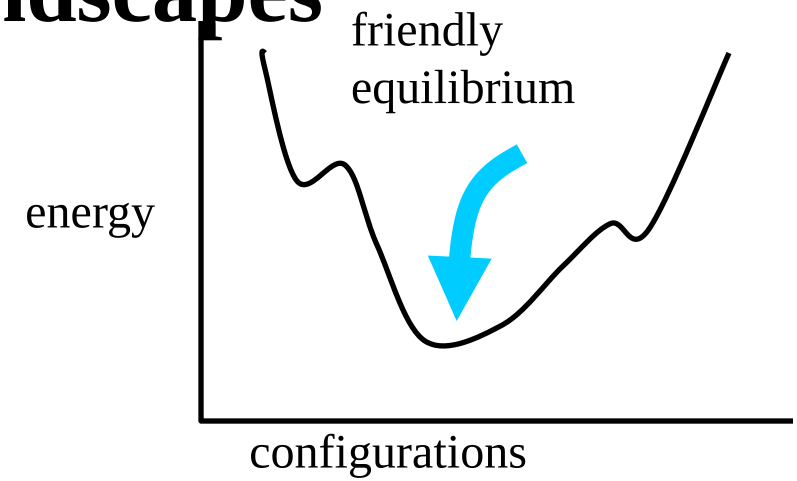
Kinetic trapping

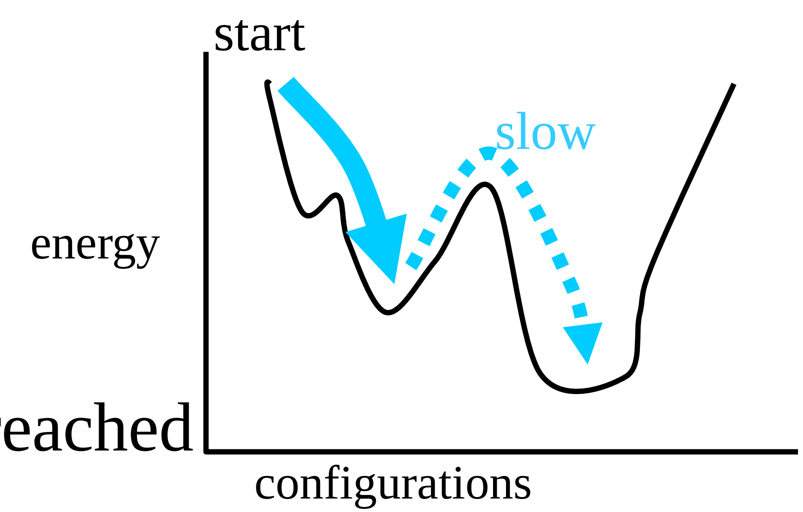
- term from protein world
- what is the friendliest energy surface ?
- wherever the molecule is
  - it will probably go to energetic minimum
- less friendly landscape



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

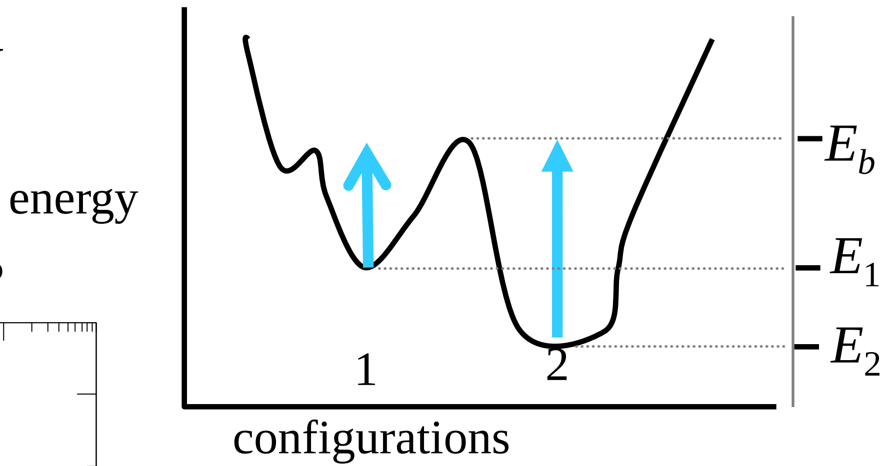
Landscapes / kinetics

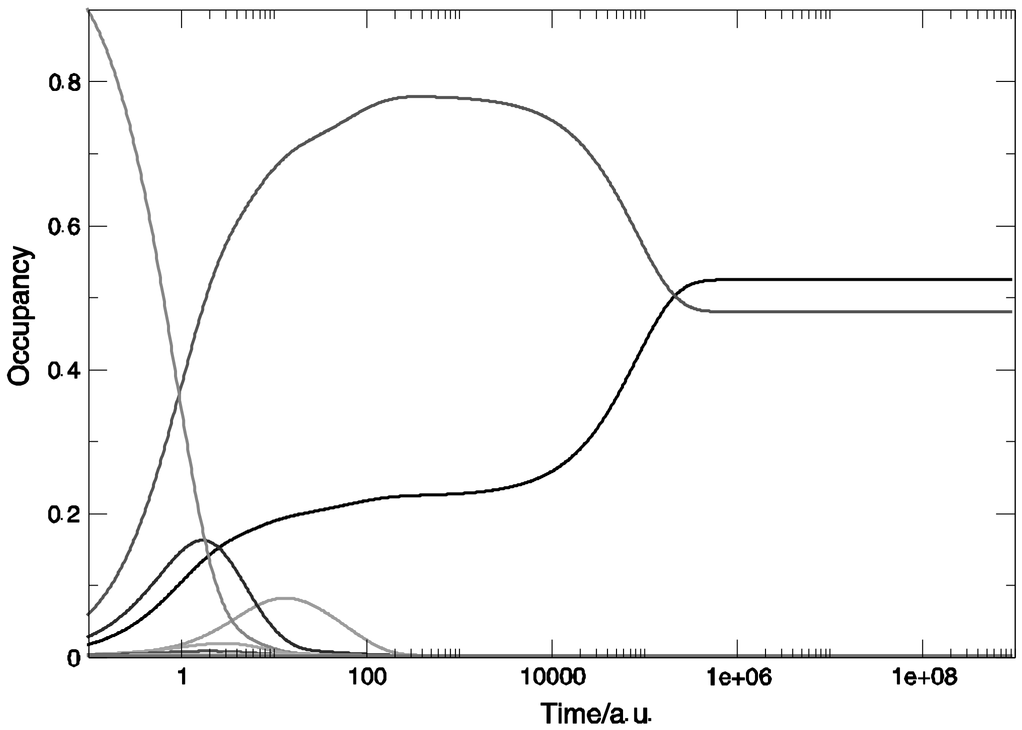
- can one predict these problems ?
  - not with methods so far
- try with simulation methods
  - Monte Carlo / time-based methods

- 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

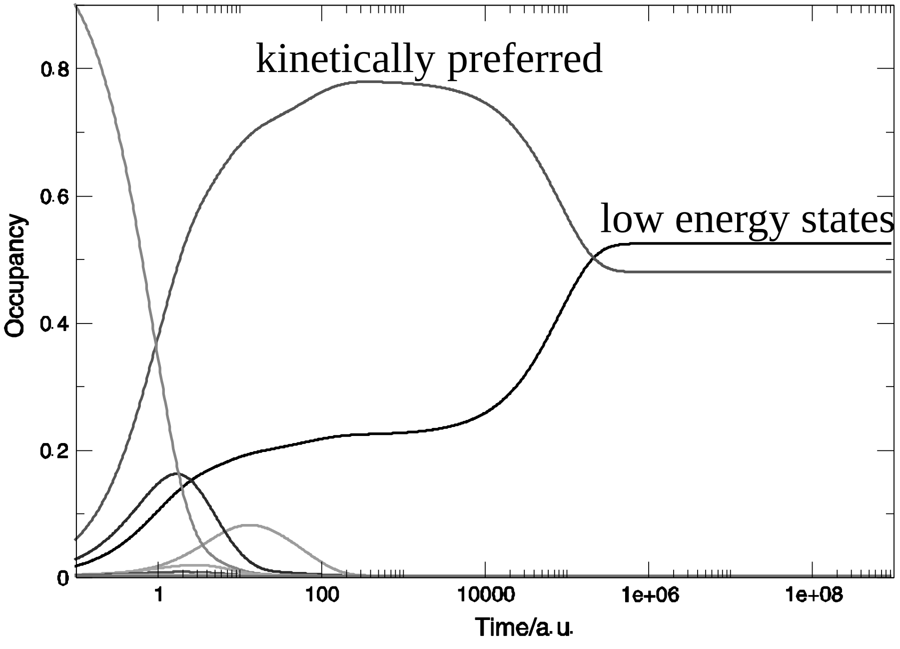




Implications

- what if RNA is degraded ?
  - molecule disappears before it finds best conformation

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



summary

- 2D (secondary structure calculations)
  - fast
    - limits structures one can predict (no pseudoknots)
  - energies not perfect
  - errors in predictions
  - may be enough for some applications where base-pairing dominates
- tertiary structure very important (binding of ligands)
- you may lose anyway (kinetics)