

Grand Plan

RNA very basic structure

very quick

3D structure

Secondary structure / predictions

The RNA world

Roles of molecules

	RNA	DNA	proteins
genetic information	yes	yes	
catalysis	yes		yes
regulation/interactions	yes	yes	yes
structure	usually single stranded	usually duplex	lots

Catalysis and binding

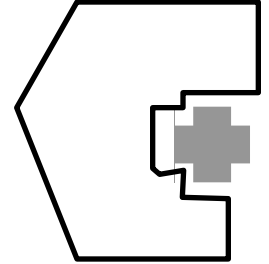
Catalysis

- proteins – classic enzymes
- RNA – less common, but well established
(ribosome, hammerhead, ..)

Specific binding

- proteins
 - bind substrates, ligands, DNA, RNA
- DNA
 - sequence specific binding – to proteins, RNA, DNA
- RNA
 - same as DNA +
 - specific catalysis implies specific recognition
 - switches and regulators

Recognition / binding specificity



Protein view – via evolution

- protein scaffold / framework positions groups
- in binding / reactive region specific groups interact
- big choice of chemical groups (20 amino acids)

DNA – not thought of in these terms

- some specificity
 - regulatory binding proteins are sequence specific
 - cleavage enzymes – sequence specific

RNA

- sequence specificity for binding proteins
- RNAszymes, aptamers, selex
- binding of arbitrary small molecules

Structure

DNA

- mostly thought of as double helix

Protein (simple dogma)

- from a specific sequence to a well defined structure
- less often – floppy, unstructured, mobile, alternative folds

RNA

- does an RNA sequence fold up to a well defined structure ?
 - all possible RNA's ?
 - biological RNA's ?
 - some RNA's ?

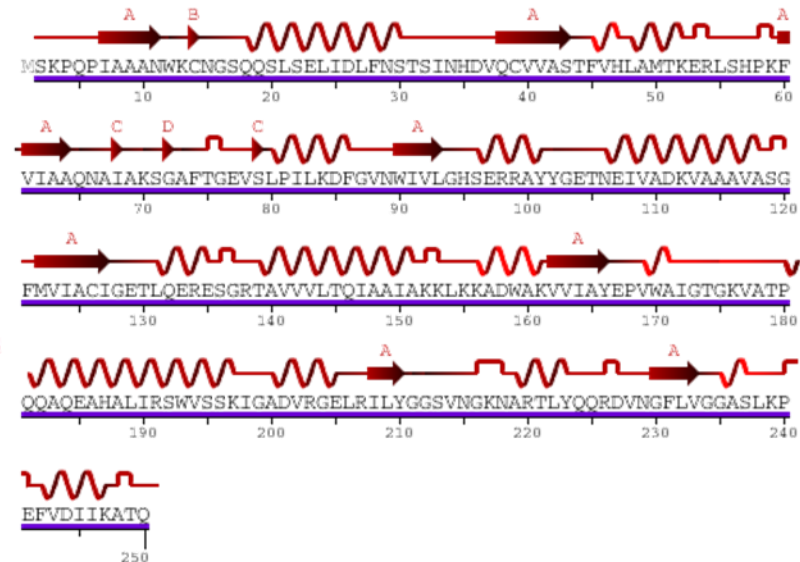
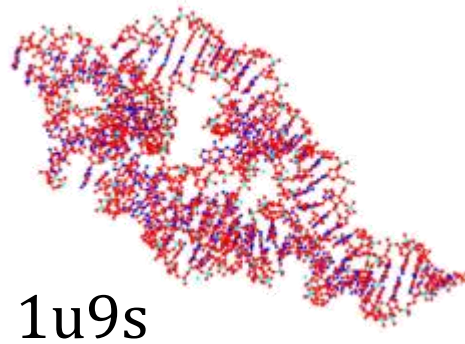
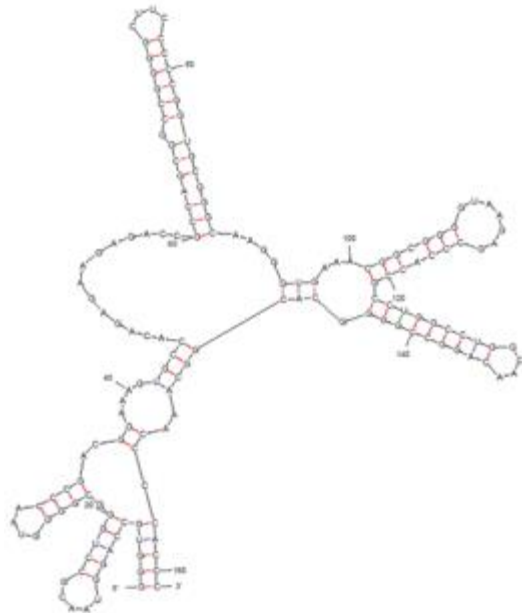
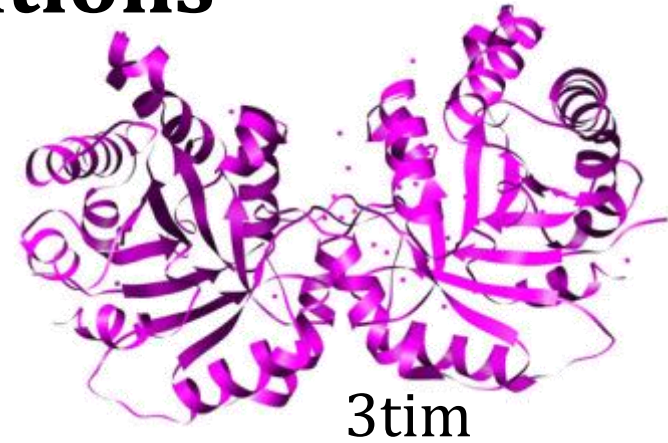
Structure Expectations

Protein

- usually 3D
- rarely secondary structure

RNA

- usually secondary structure



Structural Data

Proteins

- 1.1×10^5 or about 3×10^4 interesting ones

RNA

- 2.8×10^3 structures with some RNA
- 45 with RNA + DNA (no protein)
- 1072 with pure RNA - many small and boring

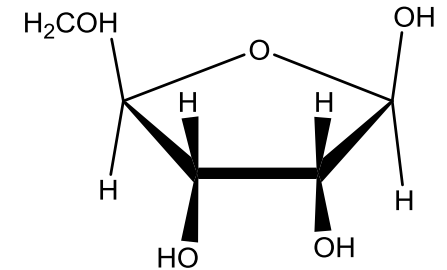
Determining structures

- general – RNA hard to handle (RNases)
- crystallography
- NMR
 - assignments very difficult (only 4 kinds of base)

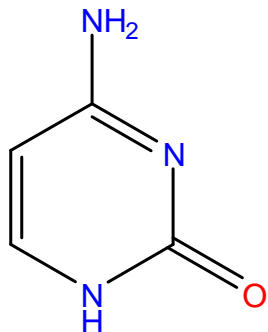
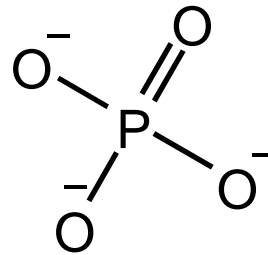
RNA structure

3 components

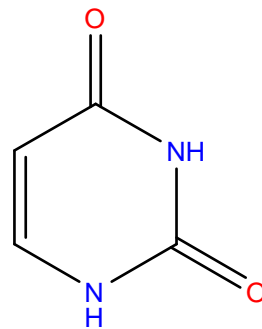
- desoxyribose (sugar)
- phosphate (PO_4)
- base (nucleotide)



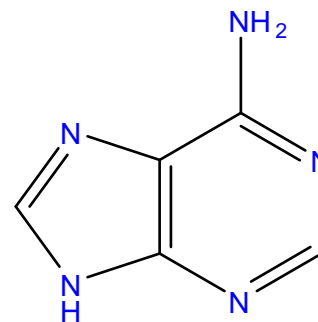
deoxy-ribose



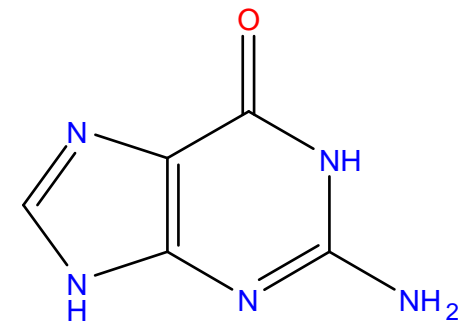
cytosine



uracil



adenine



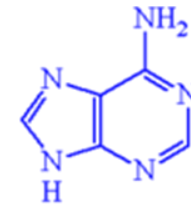
guanine

RNA Bases

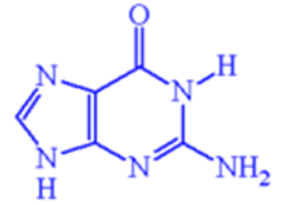
Are they like protein residues ?

- not classified by chemistry
- do they have interactions ?
 - yes

purines

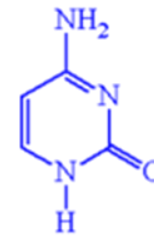


Adenine

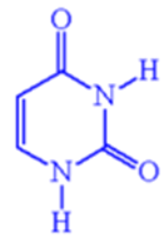


Guanine

pyrimidines

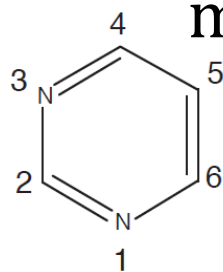


Cytosine

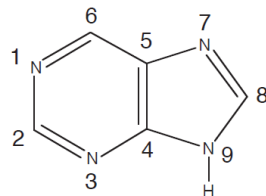


Uracil

mother shapes



pyrimidine



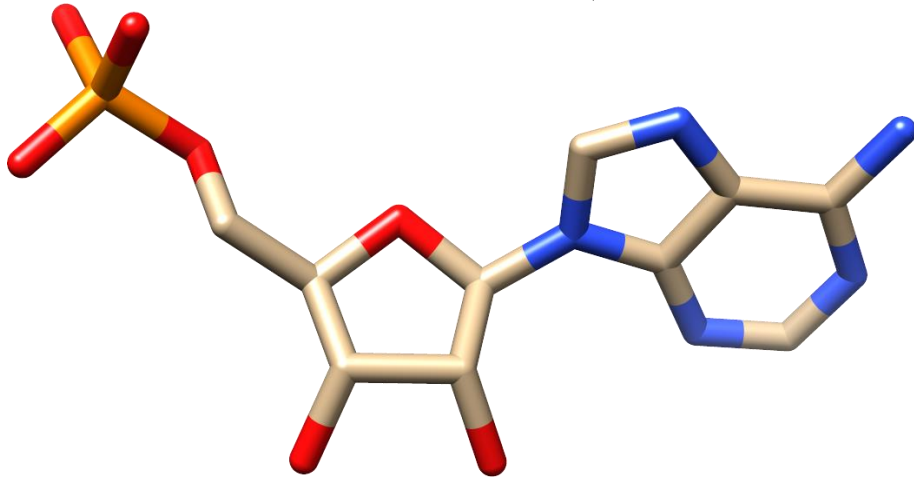
purine

- numbering not used much
- putting pieces together...

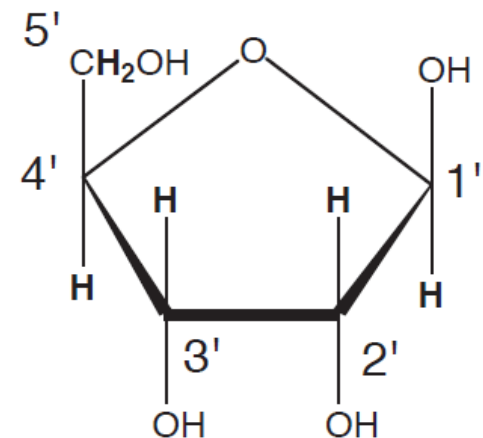
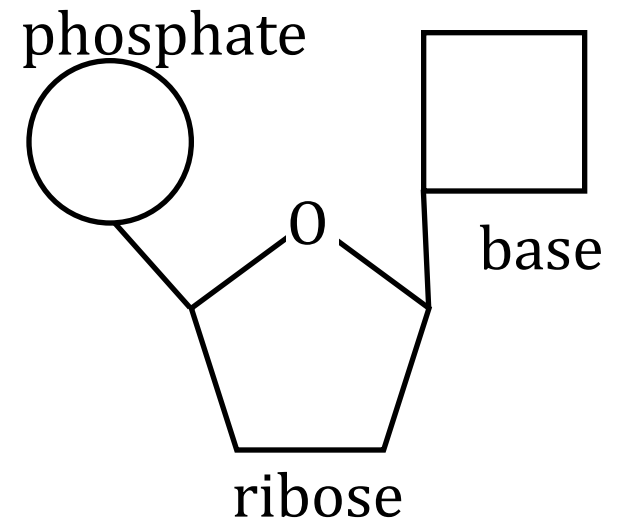
RNA structure

Joining the components

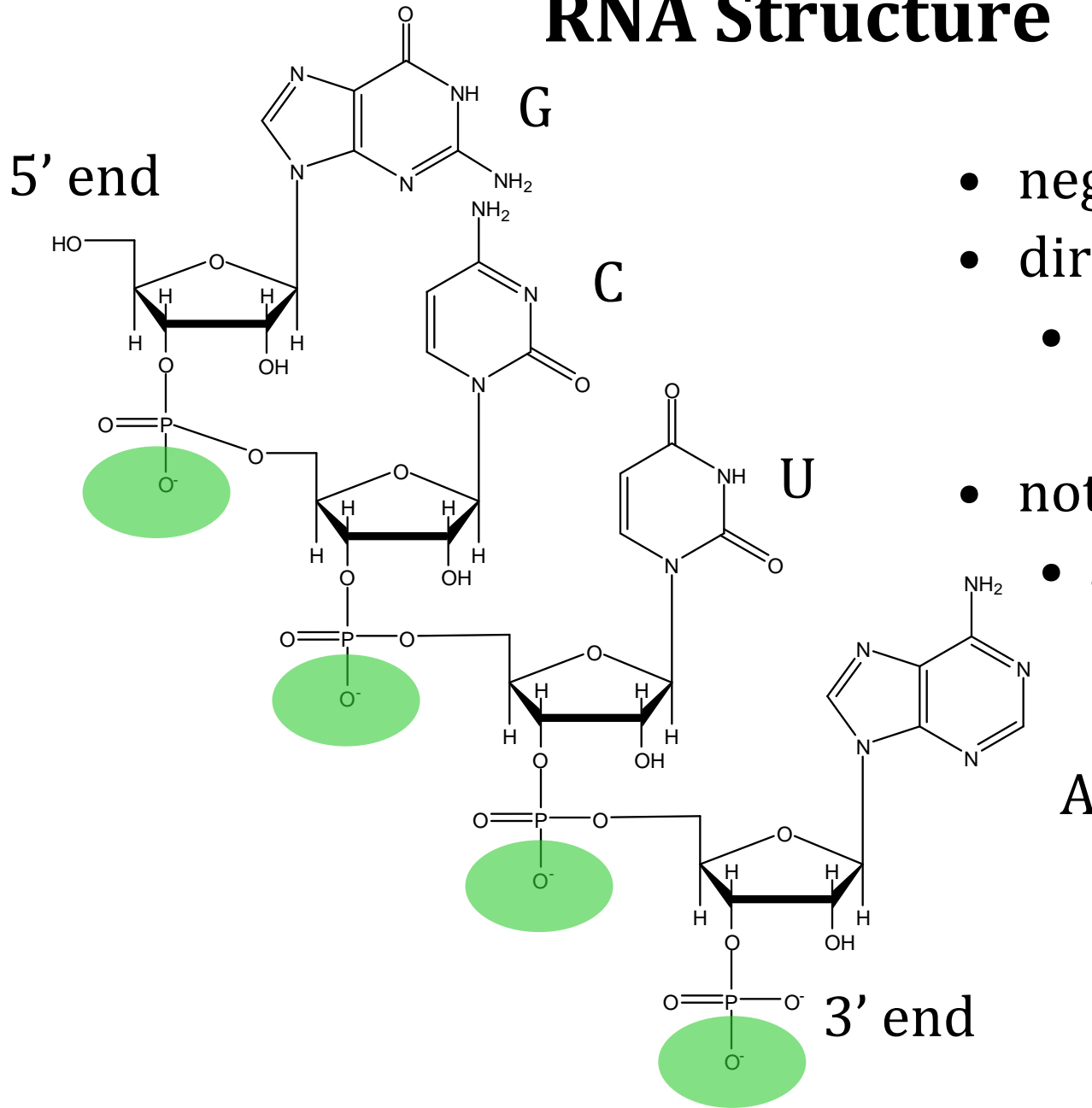
- adenosine 5'-monophosphate
 - not adenine, adenosine, ...



- note numbering on sugar ring



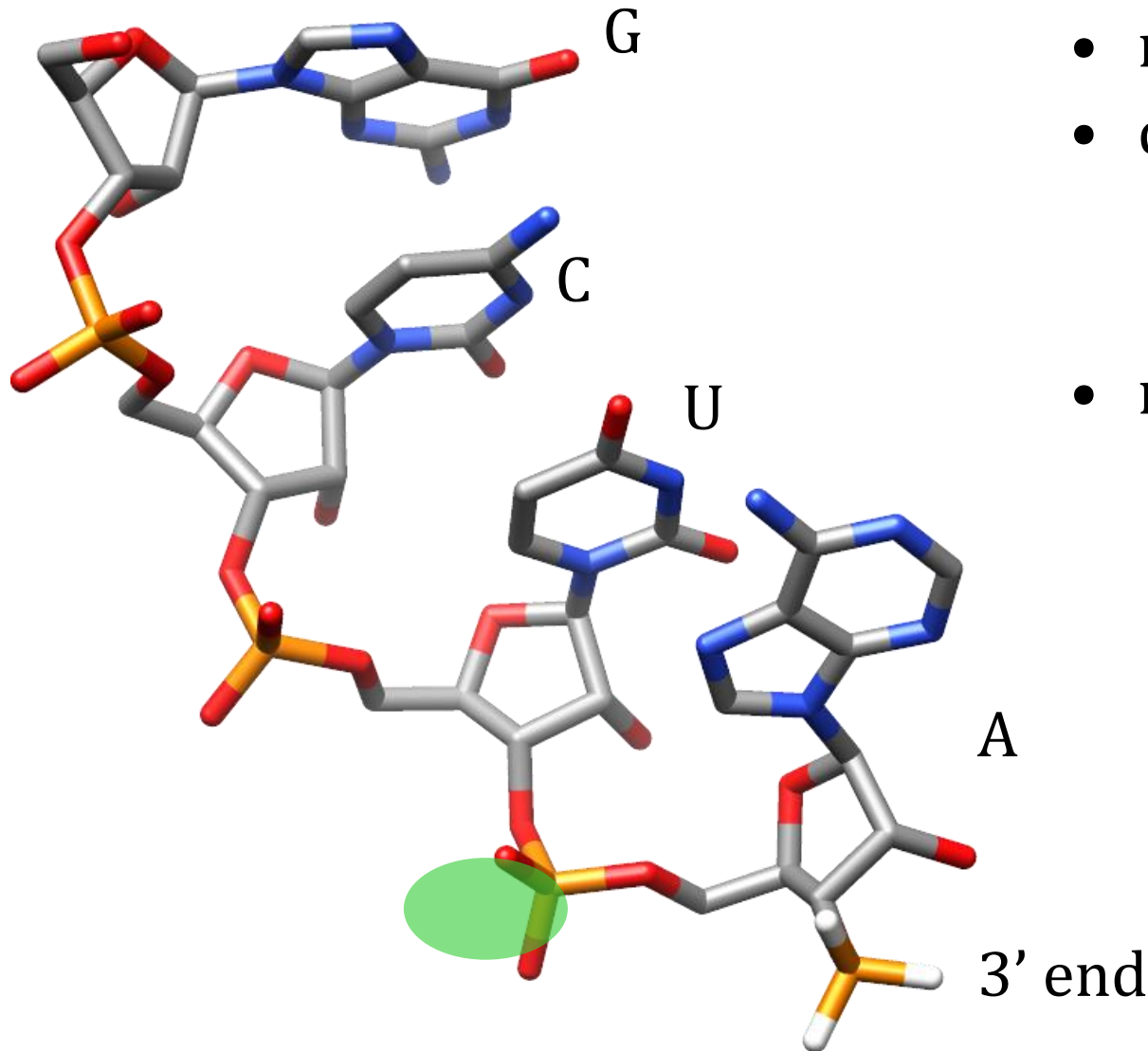
RNA Structure



- negative charges
- directional
 - 5' to 3'
- notation
 - always 5' to 3'

RNA Structure

5' end



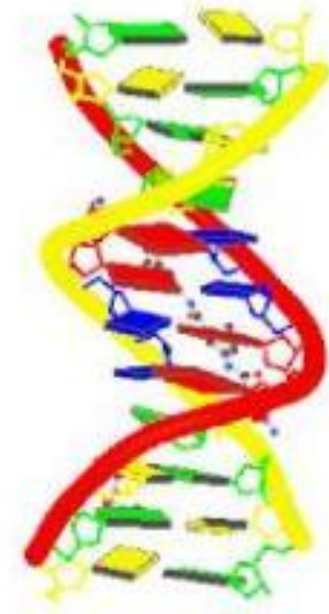
- negative charges
- directional
 - 5' to 3'
- notation
 - always 5' to 3'

H bonding

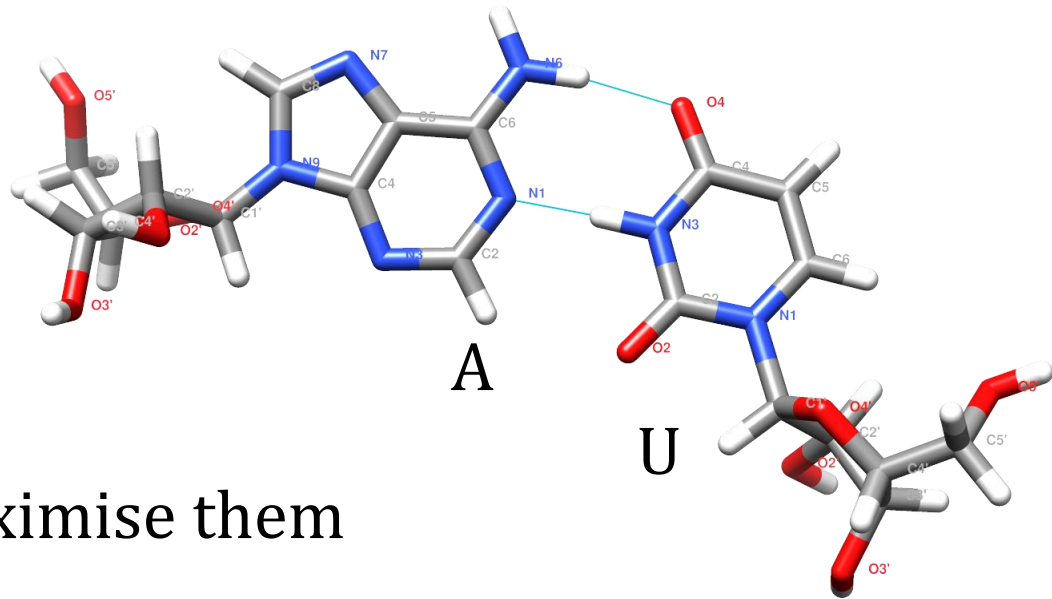
What holds the pairs of a helix together ? H-bonds

- applies to RNA
- rules from proteins
 - H-bond donors are NH, OH
 - acceptors – anything with partial -'ve

Historic H-bonding pairs...

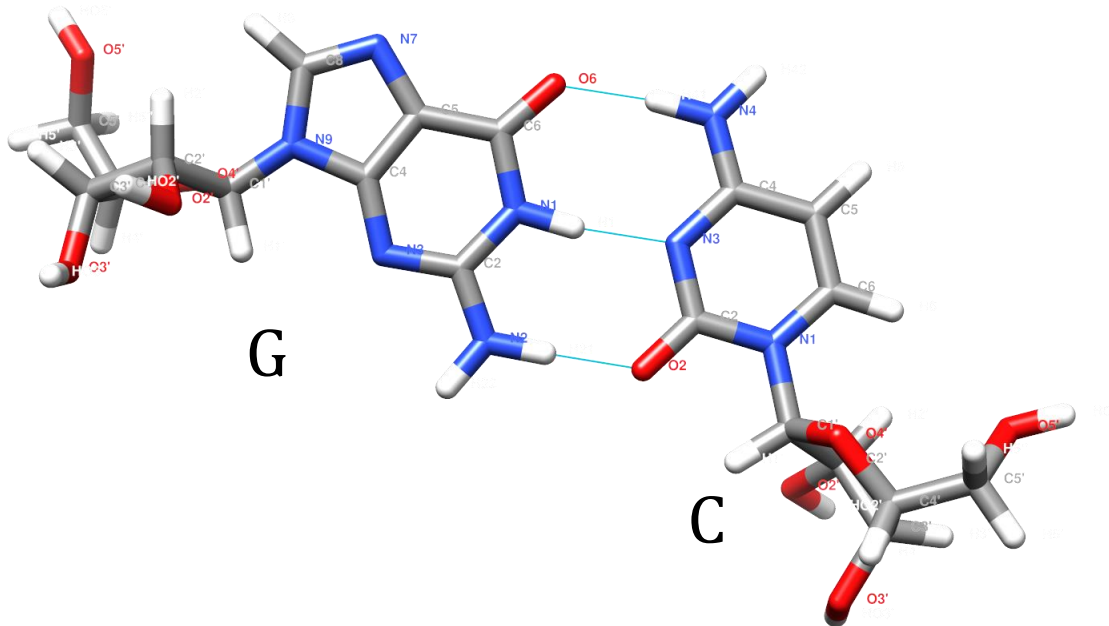


Historic H-bonding pairs



Count H bonds

Structures like to maximise them



Historic viewpoint

- RNA has 4 bases + GC, AU base pairs
- H-bond pairs look flat
 - not true

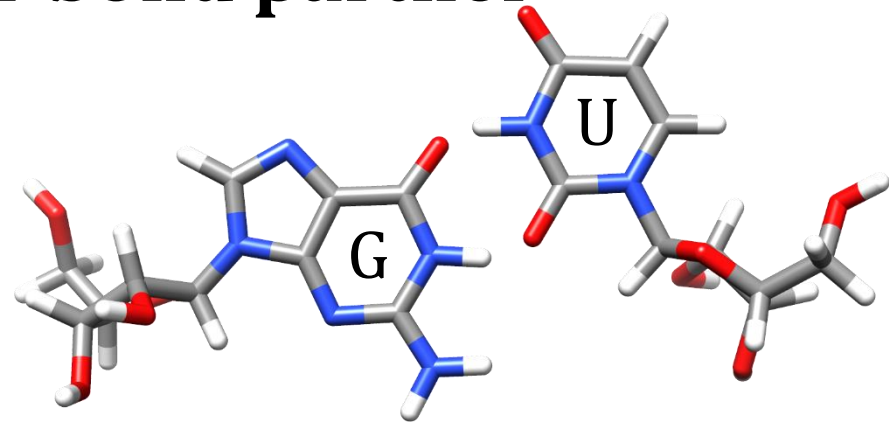
Other common H-bond partner

Contrast with DNA (GC and AT)

- almost no mismatches in DNA

RNA (GC, AU) much more interesting

- third base pair GU (rather common)
- lots of weaker pairs possible



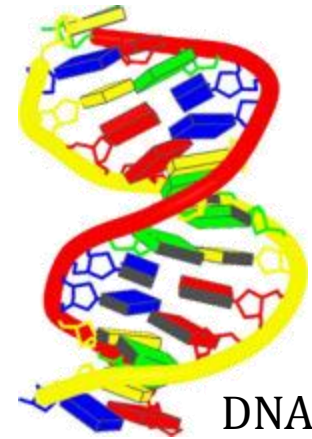
Possible RNA structures

DNA ? nearly always similar helix

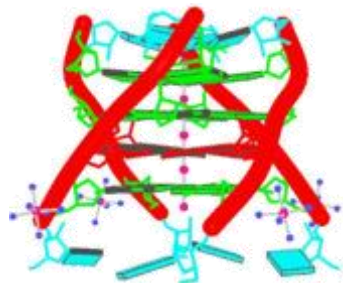
- some debate about A, B, Z, ..

RNA

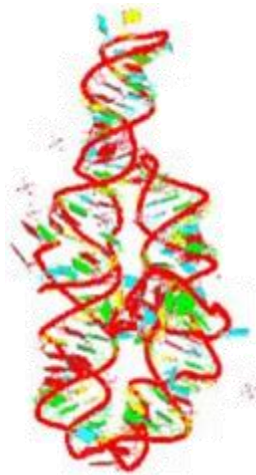
- lots of varieties known
- nomenclature..



DNA
duplex
140D



tetraplex
1mdg



group I intron
1hr2



hammerhead
2oeu



tRNA
1evv

RNA coordinates / nomenclature

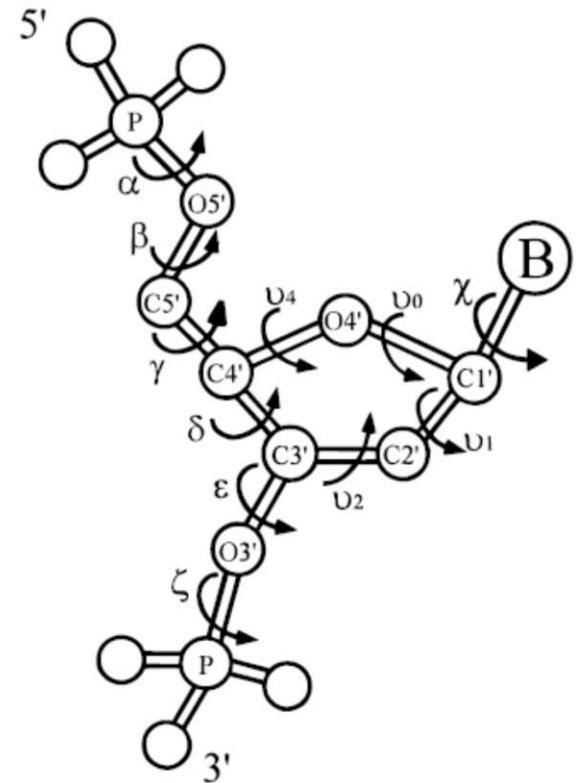
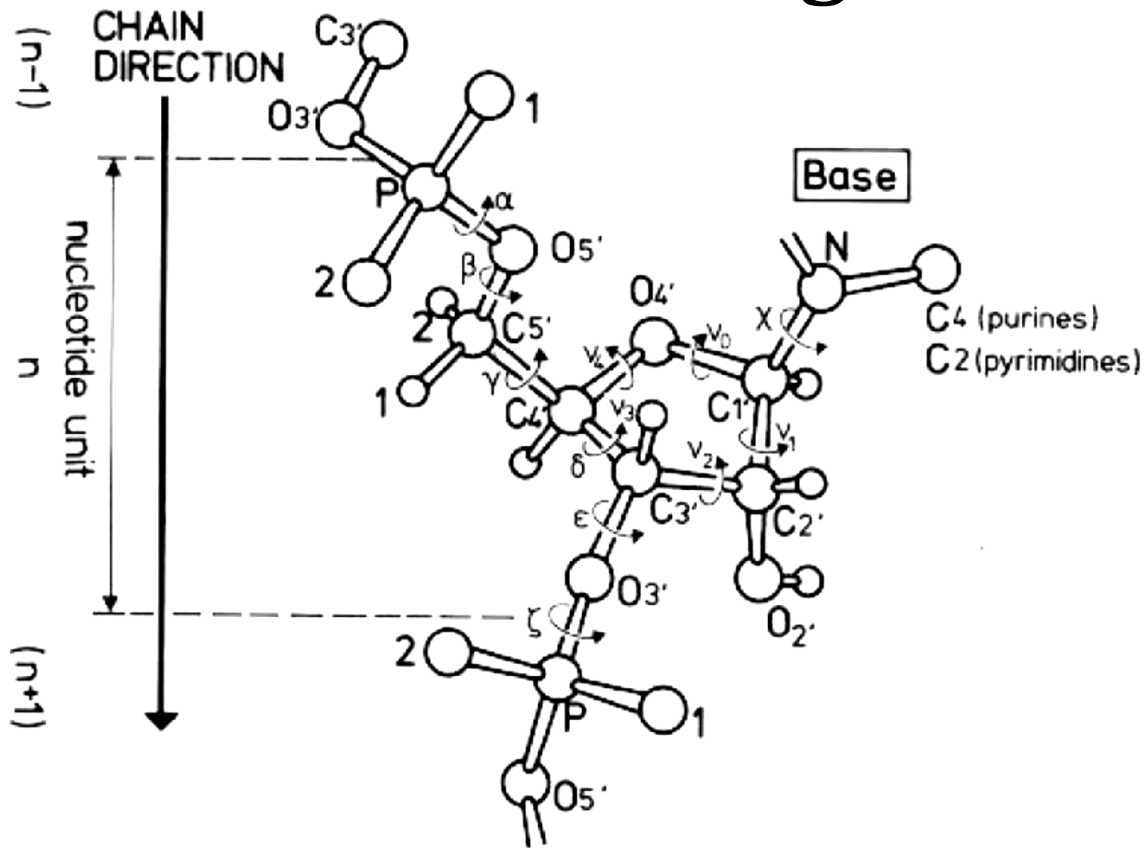
As for proteins: PDB format

ATOM	1	O5*	G A 103	58.355	47.332	91.116	1.00175.32
ATOM	2	C5*	G A 103	57.373	48.210	90.636	1.00175.32
ATOM	3	C4*	G A 103	56.962	47.802	89.224	1.00175.19
ATOM	4	O4*	G A 103	58.148	47.463	88.474	1.00175.34
ATOM	5	C3*	G A 103	56.096	46.543	89.152	1.00175.03

As for proteins

- dihedral angles are useful
- Unlike proteins (φ, ψ) there are 8 ($\alpha, \beta, \gamma \dots$)

dihedral angle nomenclature



from Marino, JP, Schwalbe, H., Griesinger, C, Acc. Chem. Res. 32, 614-623 (1999)

dihedral angle nomenclature

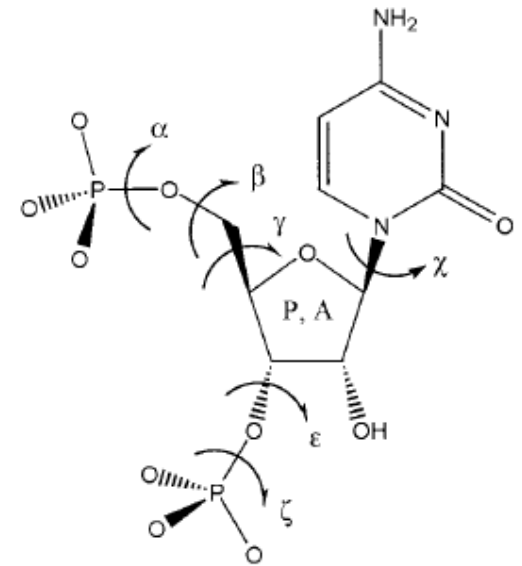
8 angles

- α , β , γ , ε , ζ , χ
- 2 for sugar (P, A)
- too many for me – how to simplify ?

what if two angles are highly correlated ?

- if we know x , then y is probably known

ideas for classification...



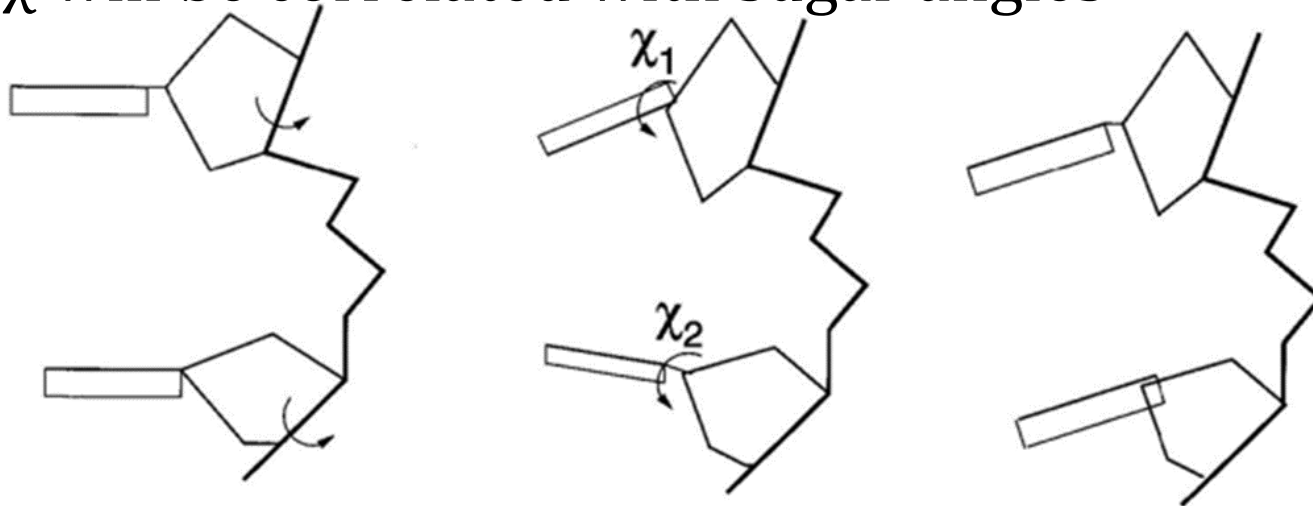
Describing RNA conformation

Example approach – look for correlations

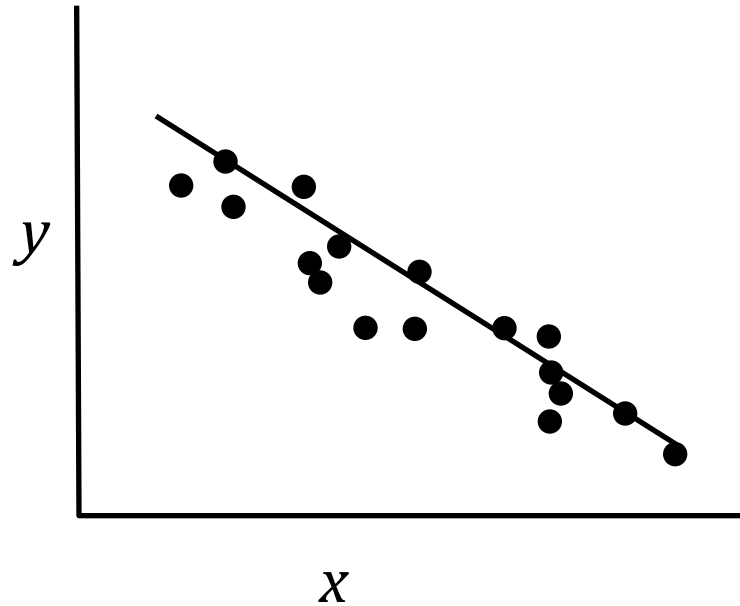
- principle component analysis (quick detour if necessary)

What if sugars move in two residues ?

- energetically, would like to maintain base pairing...
- P, A move, χ will compensate
 - χ will be correlated with sugar angles



PCA reminder



I have two dimensional data

- could well be described by a first (component) and
- maybe second component

n -dimensional data

- how much of variance is described by 1st, 2nd, ... components

Describing RNA structure

- Collect data for all angles
- Use principle component analysis to see what is important

Claim

- conformations are well described by just 3 angles

An alternative

- do not think in terms of classic angles

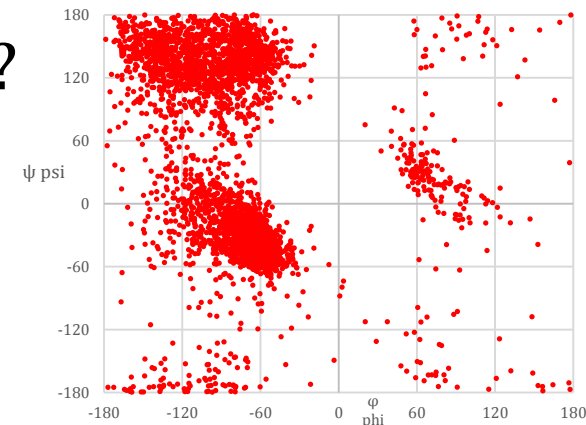
Describing RNA conformation

Alternative...

- do not work in terms of real dihedral angles
- invent reference points
- example study...
 - Duarte, CM & Pyle, AM, (1998) 284, 1465-1478

remember ramachandran plots in proteins

- can one do something similar in RNA ?

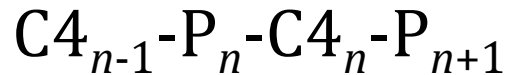


Reduced RNA conformation

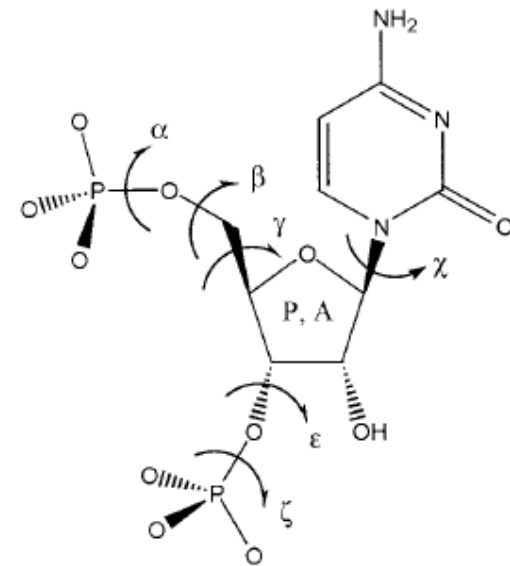
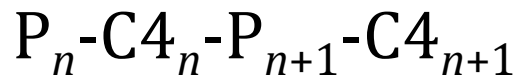
Basic idea

- pick 4 atoms that are not sequential
- define a simplified backbone
 - P-C₄-P-C₄-P-C₄-...
- leads to "pseudo-torsion" angles

η



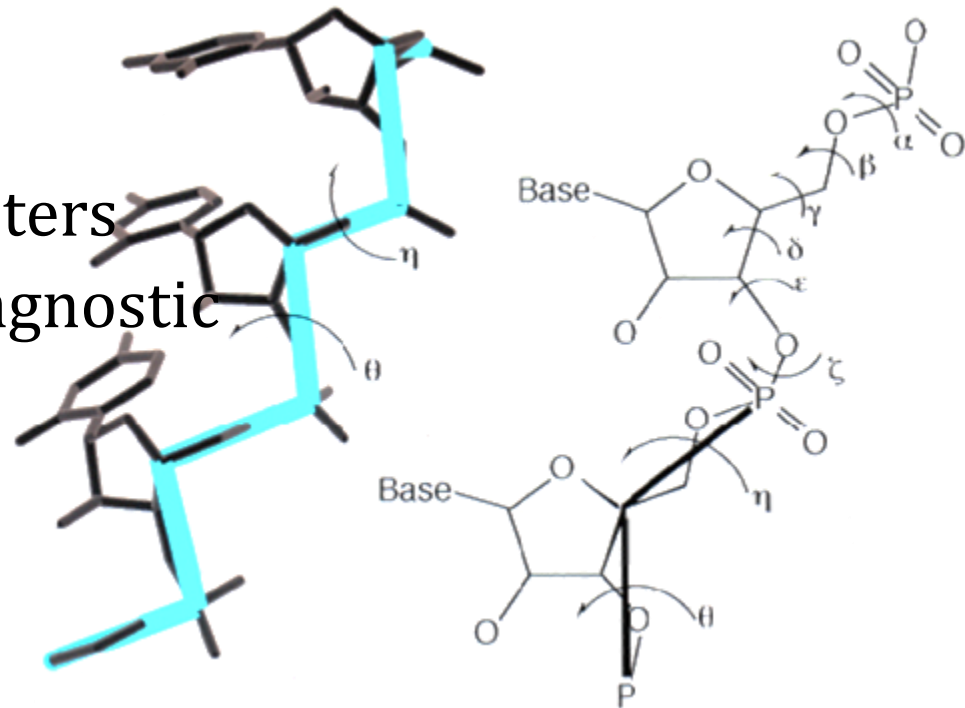
θ



Reduced RNA conformation

Plan of authors

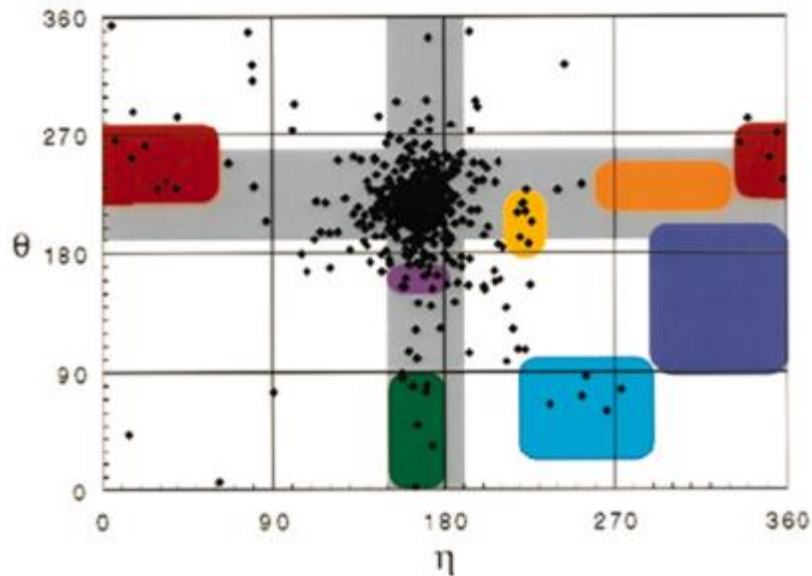
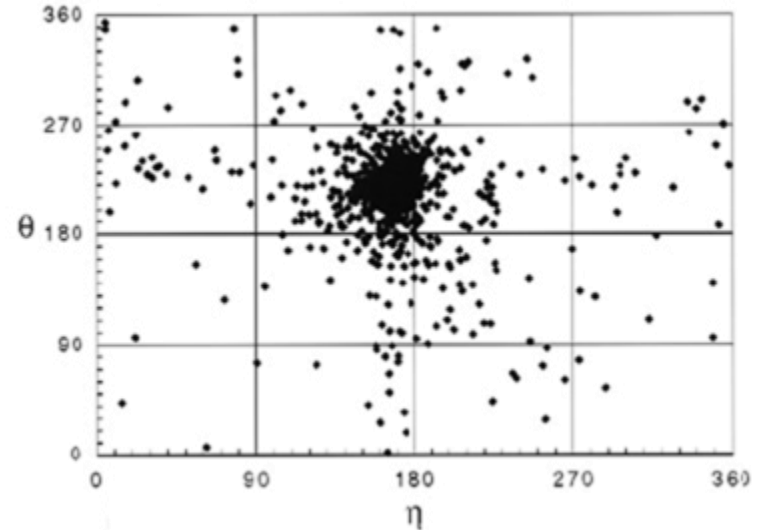
- take 52 structures
 - (≈ 700 nucleotides)
 - collect η, θ
 - see if there are clusters
 - see if angles are diagnostic



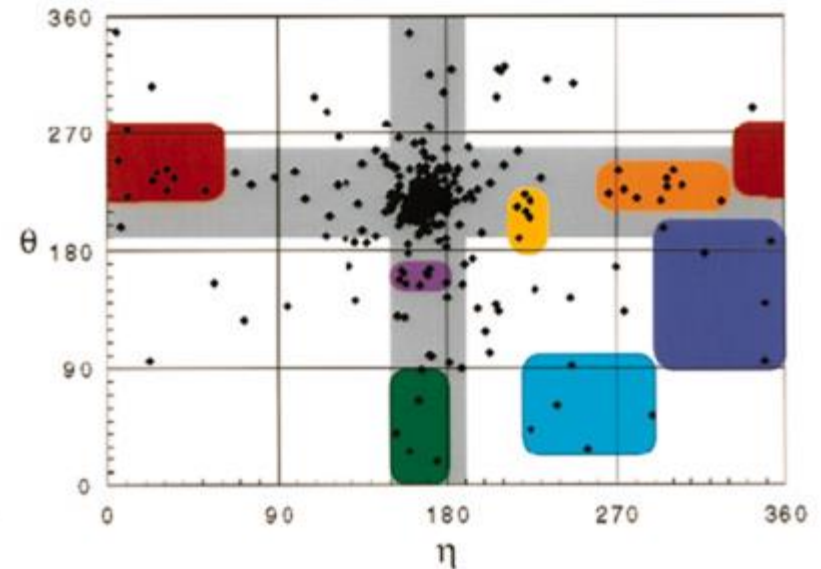
Reduced RNA conformation

Do you see clusters ?

- main set of points ...
- boring RNA helix
- a big claim

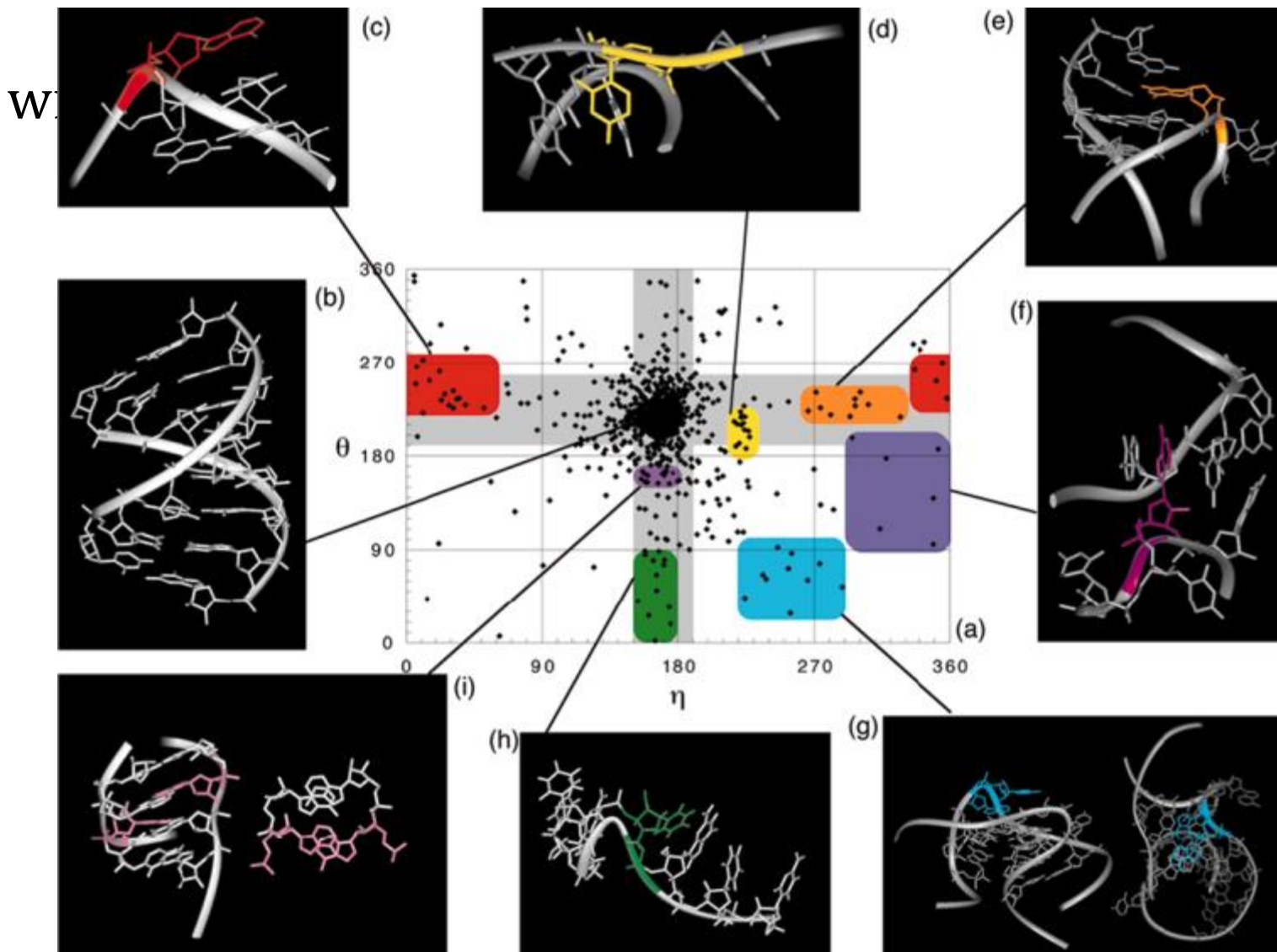


no tertiary interactions



yes tertiary interactions

Reduced RNA conformation



Reduced RNA conformation

We are interested in a critical look at ideas

How to read this...

- if you measure a pair of η, θ pseudo-angles
 - could you guess if something is wrong in structure ?
 - could you use this to categorise the conformation ?
- are there better ways to categorise structure ?

Summary

- RNA structure as per Watson-Crick, old text books
- How are RNA structures different to DNA ?
- What are the biological roles ?
- Can we neatly summarise RNA structures ?
 - see what information (angles) are necessary
 - define alternative angles
- Next..
 - predicting secondary structure

RNA structure, predictions

Themes

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

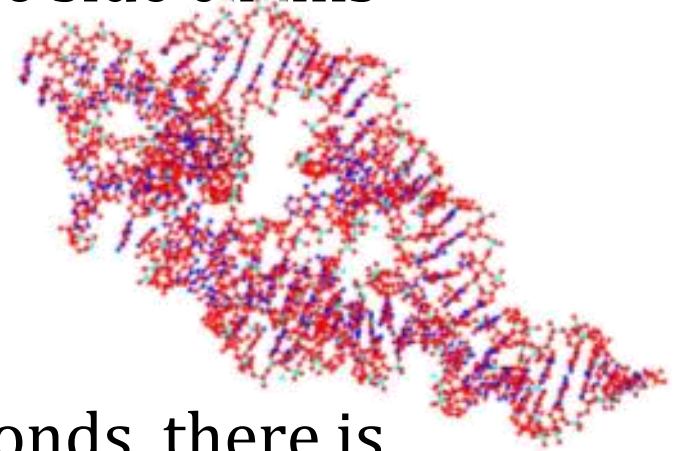
Structure – protein vs RNA

Middle of proteins

- hydrophobic core - soup of insoluble side chains

Middle of RNA

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



Protein structure lectures are not helpful today

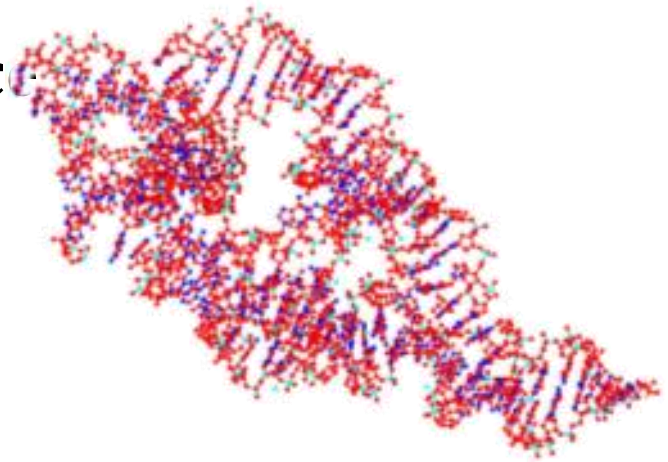
RNA - how important is 3D structure ?

Binding of ligands (riboswitches, ribozymes)

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

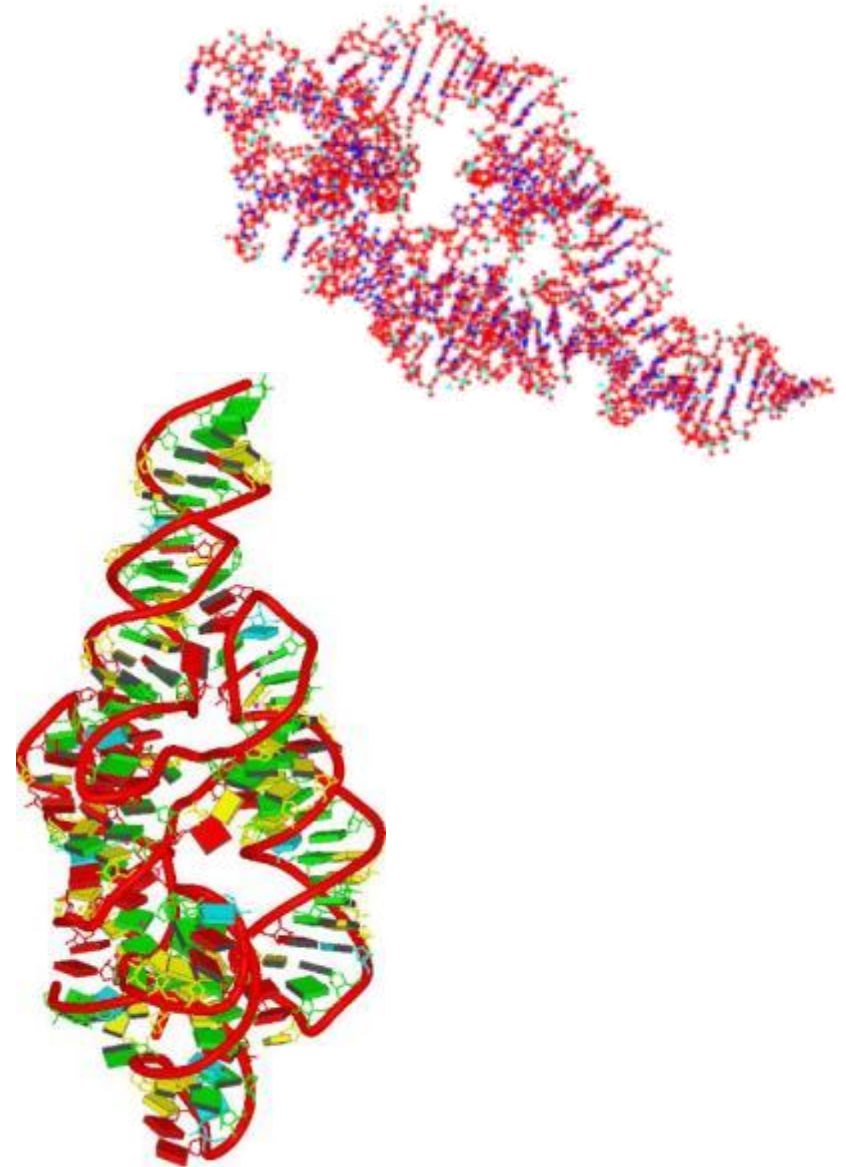
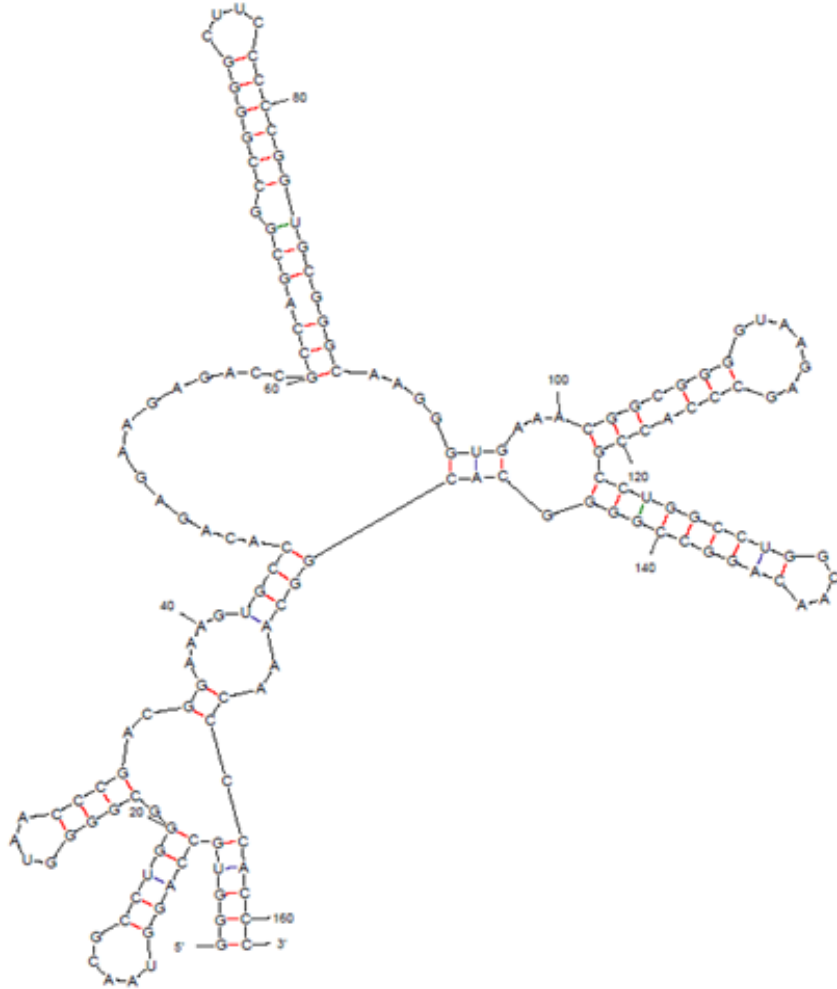
What do we do ?

- mostly ignore it



How realistic is 2D ? How relevant ?

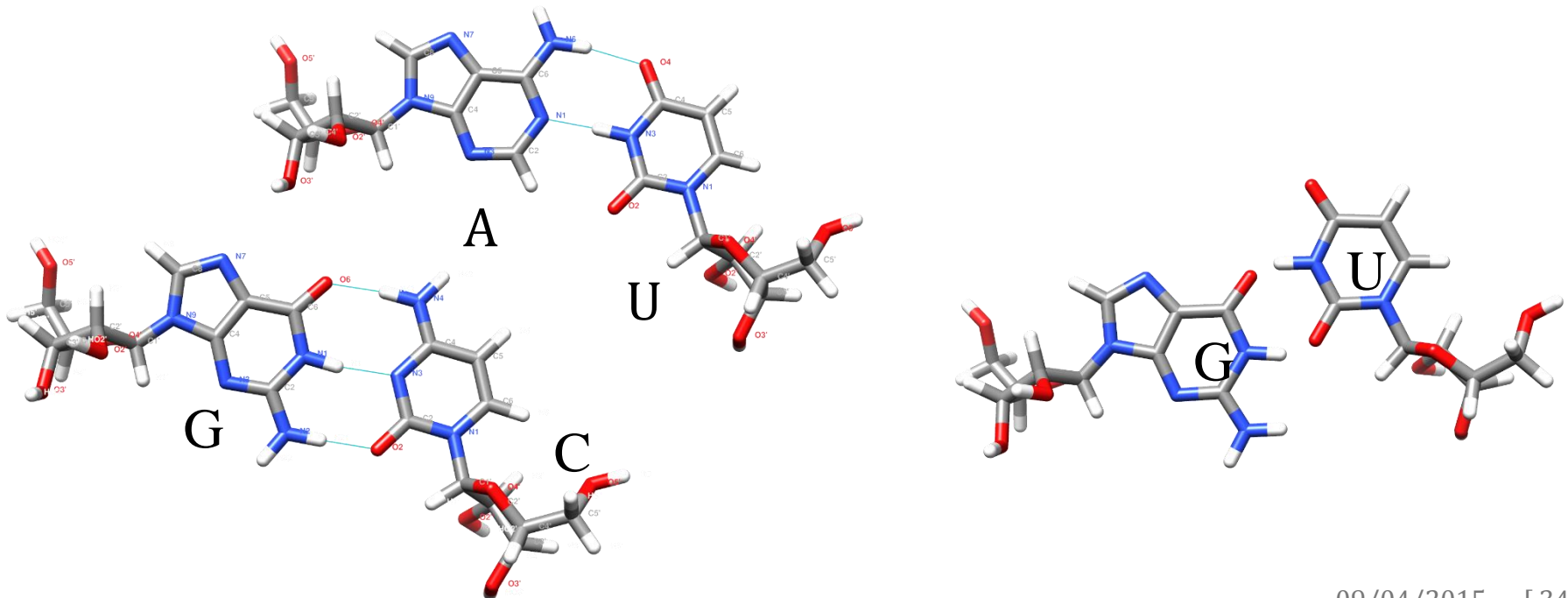
3D versus 2D



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)

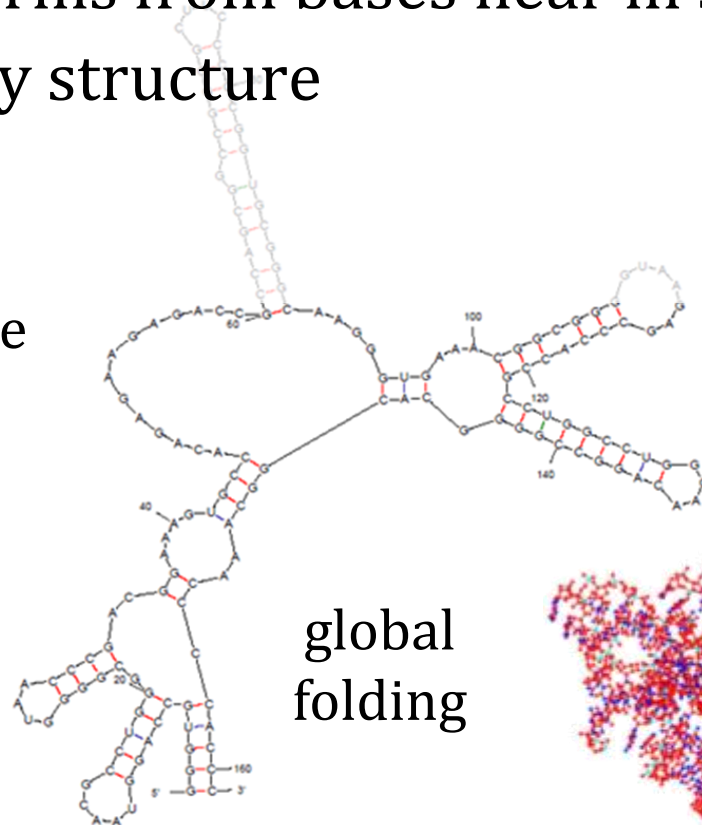


2D why of interest ?

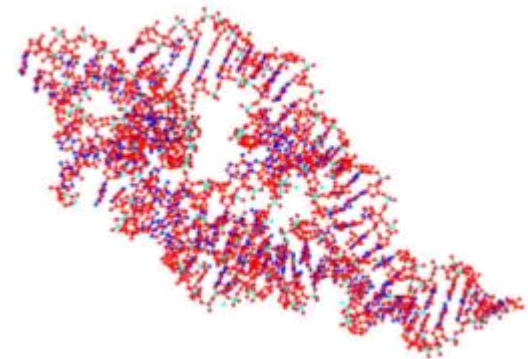
3. Claim - RNA folds hierarchically

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

secondary structure



global
folding



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 ?

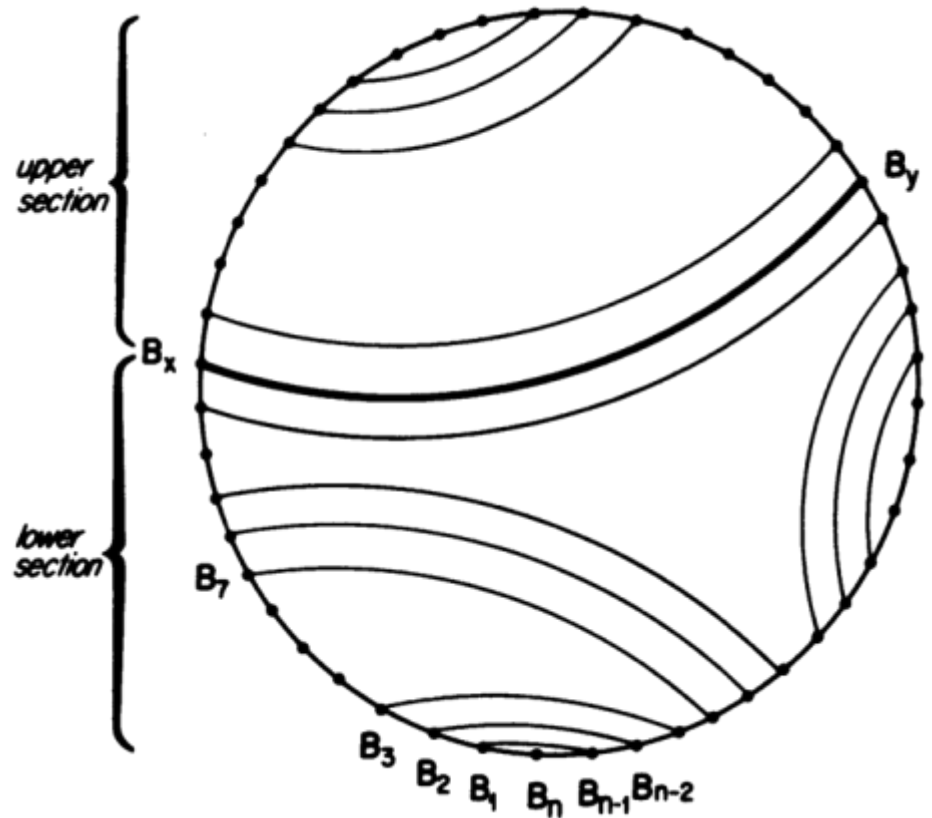
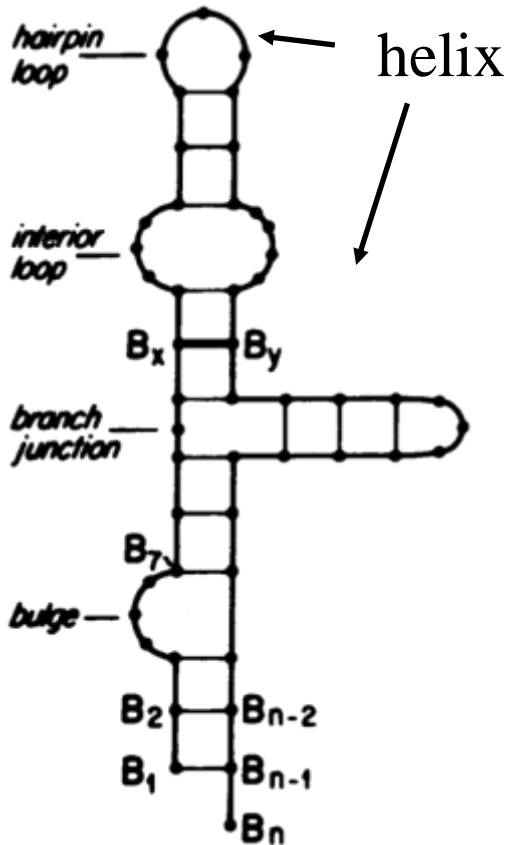
- RNA double strand helices are believed to be stable

Useful ? if true

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

Four representations of flat RNA

1. conventional

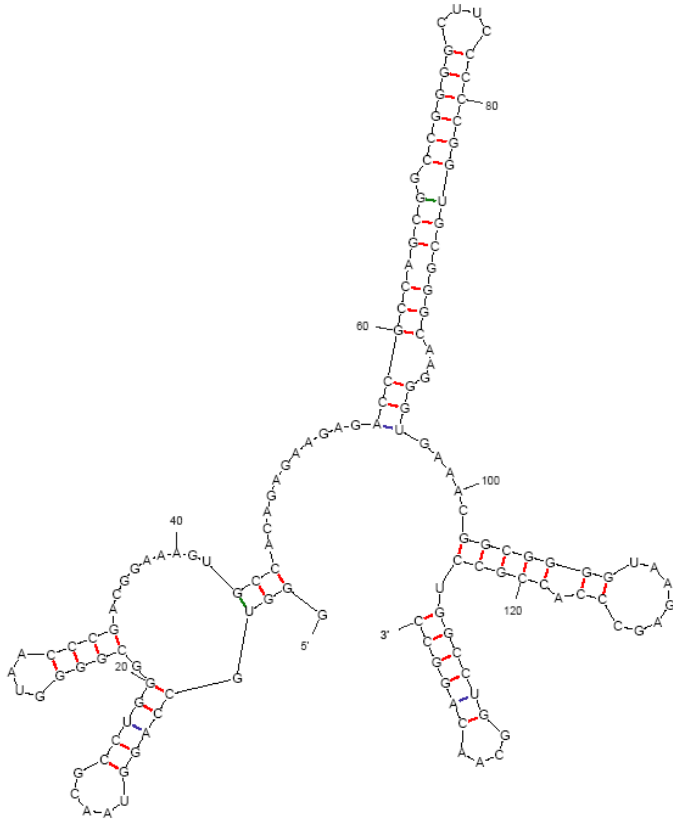


2. Nussinov's

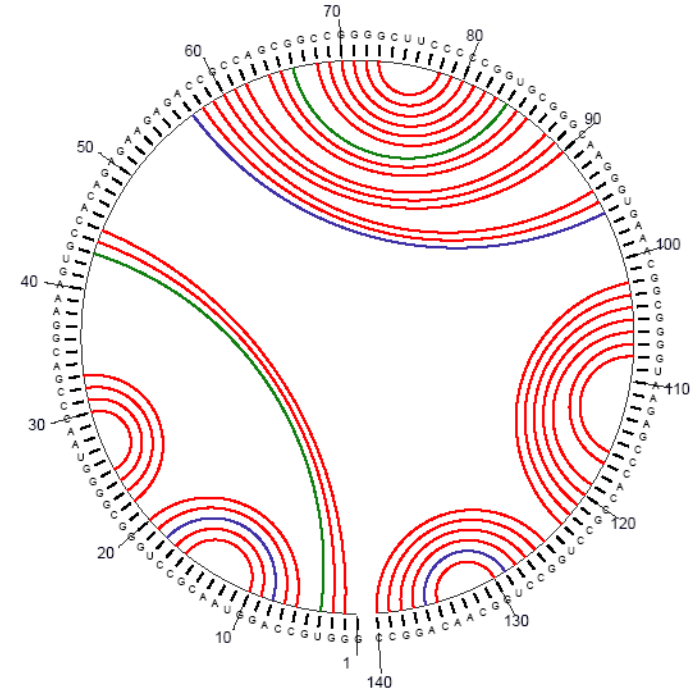
- write down bases on circle
- arcs (lines) may not cross

+ on next slide

Four representations of flat RNA



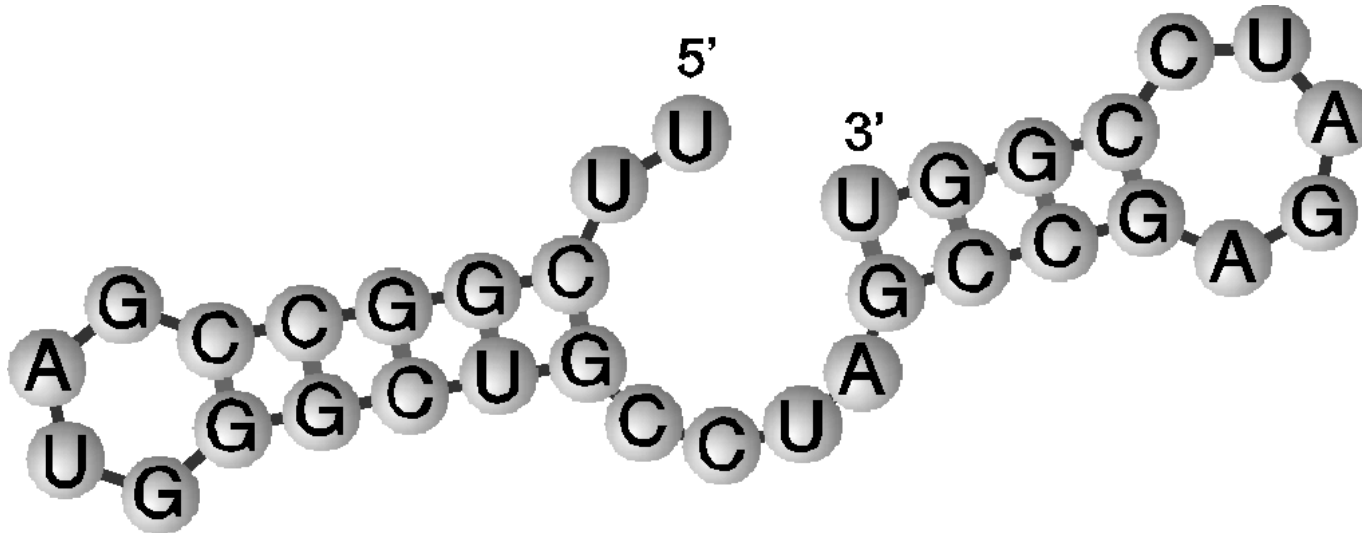
1. conventional representation



2. Nussinov's circle

Same features on both plots

Parentheses

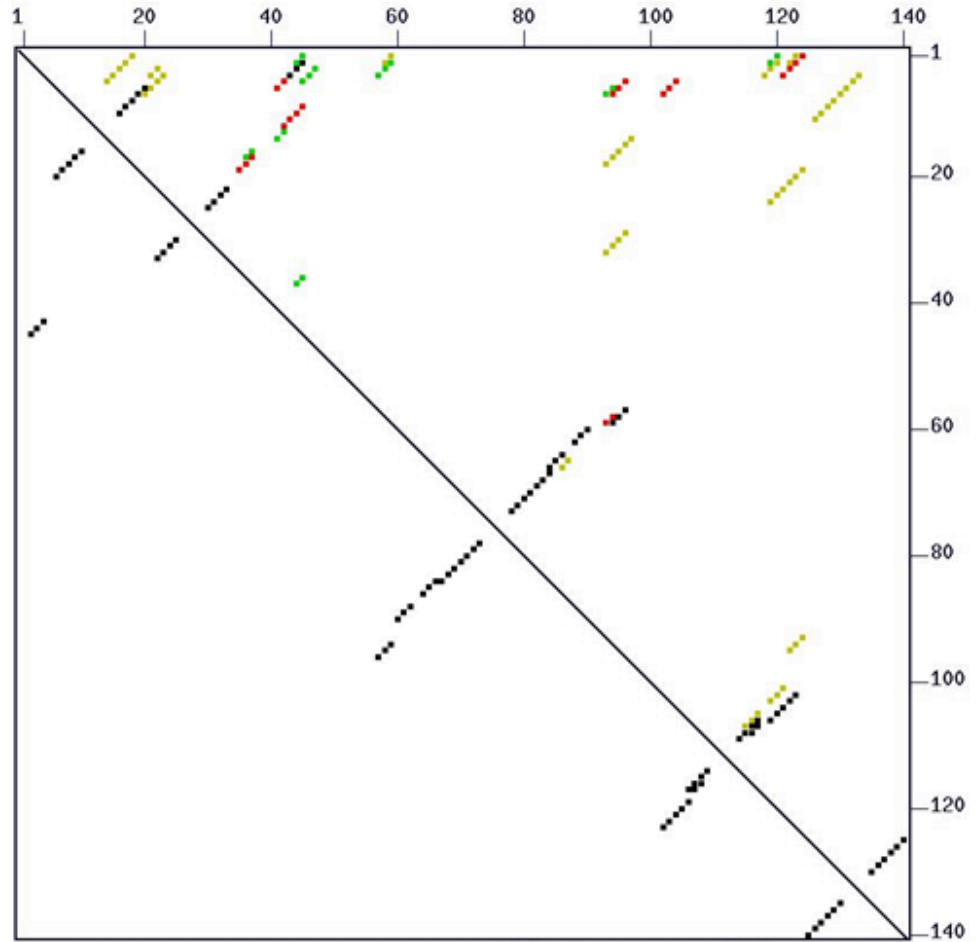
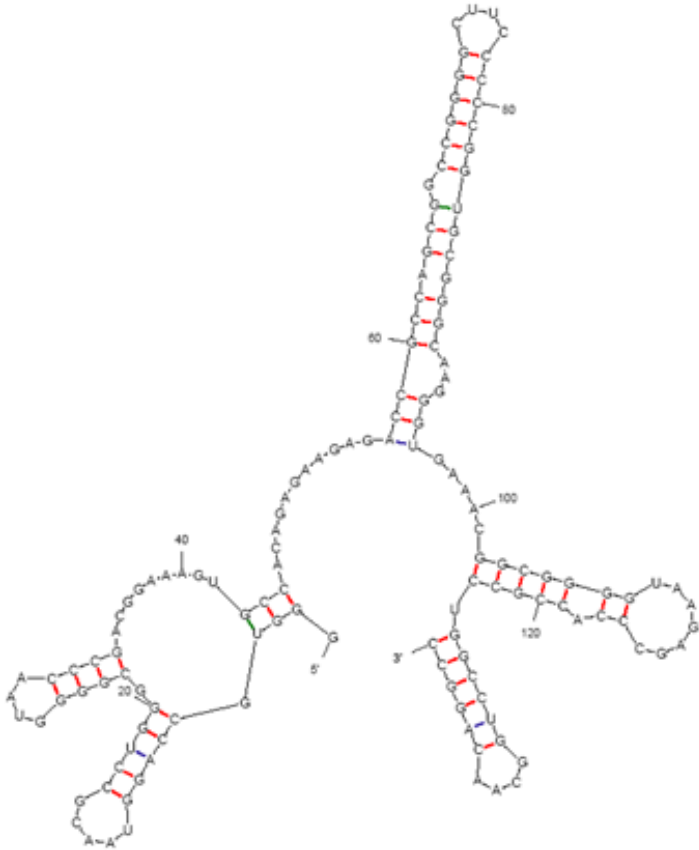


3. parentheses – most concise

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

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

Dot plots

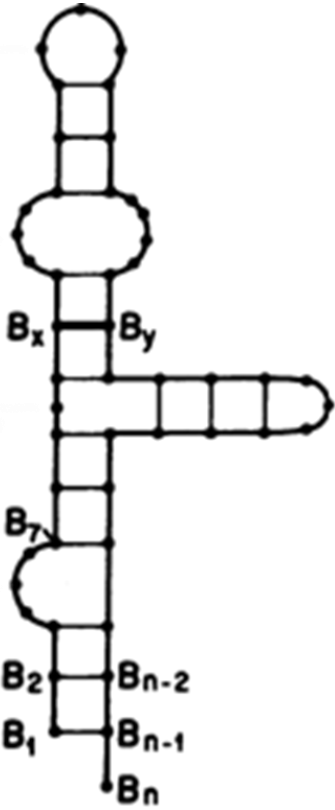


4. Dot plots

Same features in both plots

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

nomenclature / features



single strand



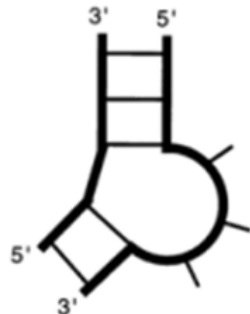
A-form double helix



Double helix with 5'-dangling end



single nucleotide bulge



three nucleotide bulge



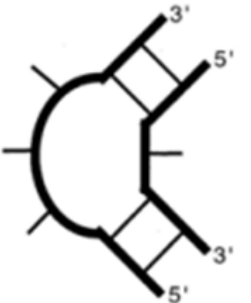
hairpin loop



mismatch pair
or, symmetric internal
loop of 2 nucleotides



symmetric internal loop



asymmetric internal loop

For explanations later

- hairpin loop
- bulge (unpaired bases)

Nussinov, R., Jacobson, A.B. Proc. Natl. 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

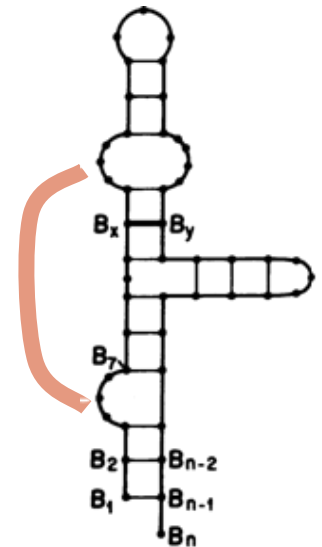
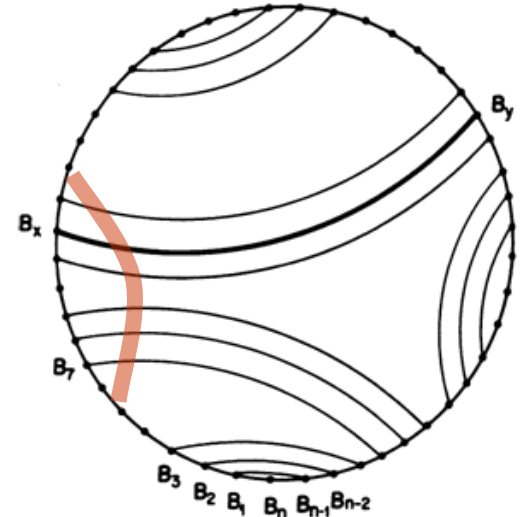
Declare crossing base pairs illegal

- 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 ?

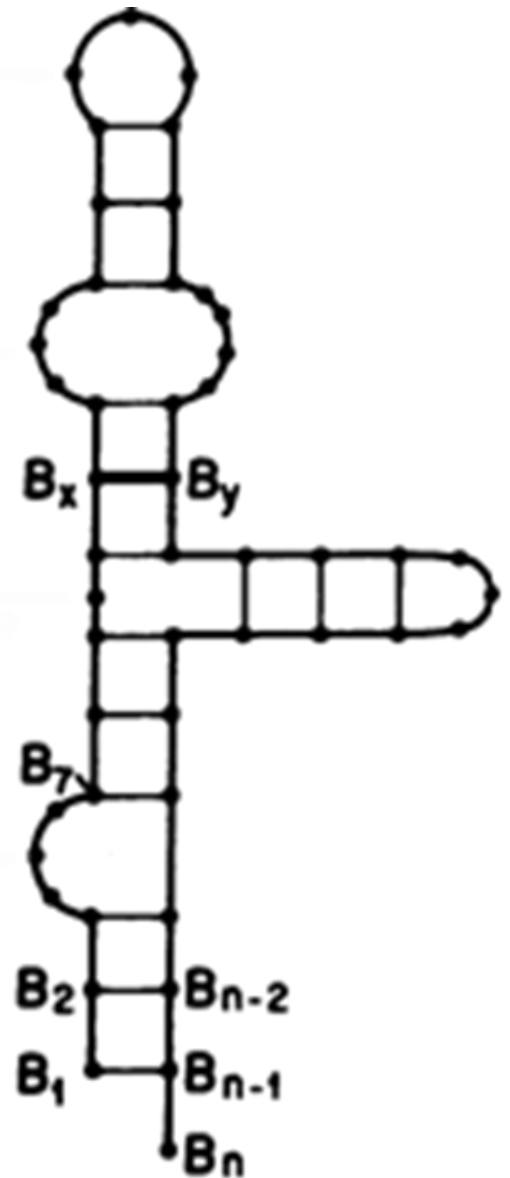
Philosophy

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

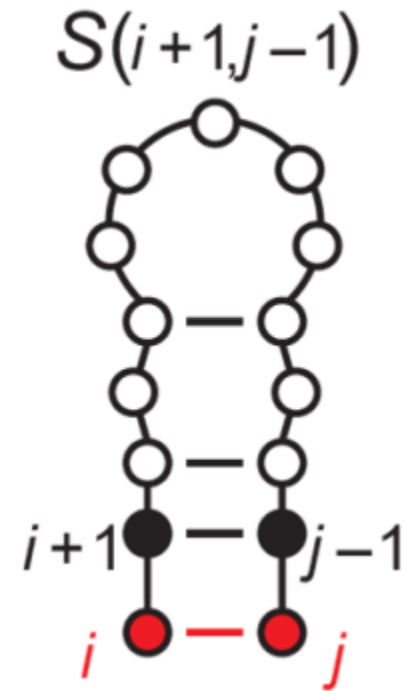
Start by looking for best possible hairpin

If we know the structure of the inner loop

- we can work out the next

If we know the black parts

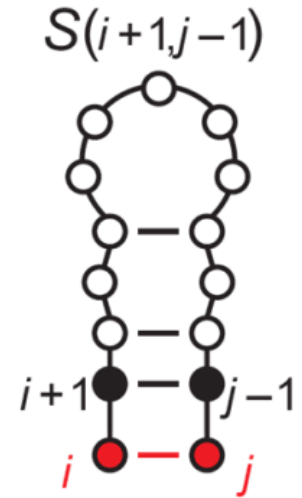
- we can decide what to do with the red *i* and *j*



hairpins / loops

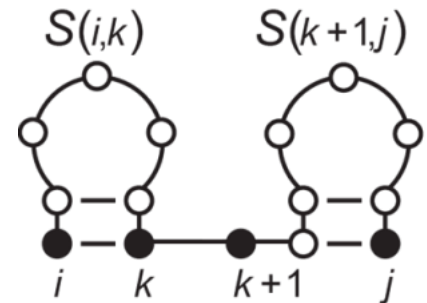
Important idea

- 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...

Pseudoknots

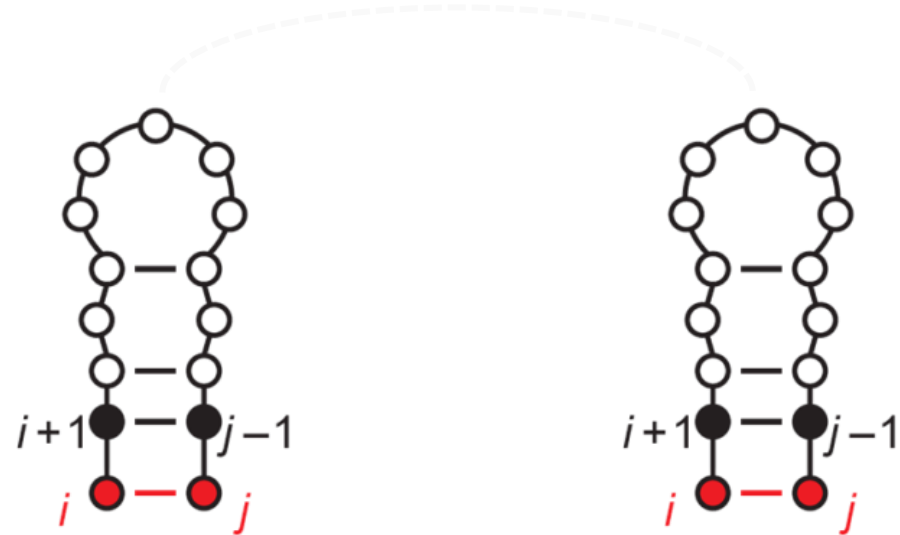
Have we considered .. ?

No !

Name – pseudoknot

Do we worry ?

- Stellingen – no
- here ? Probably.



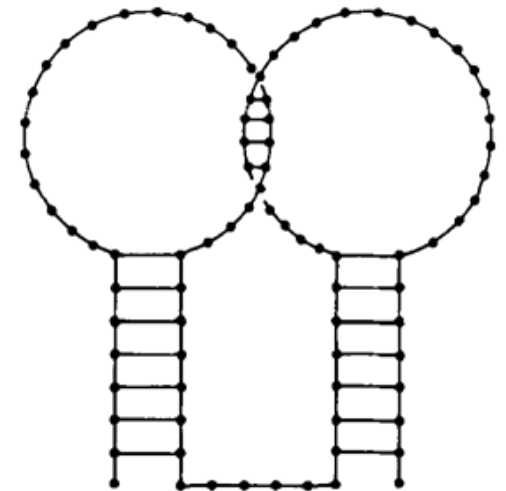
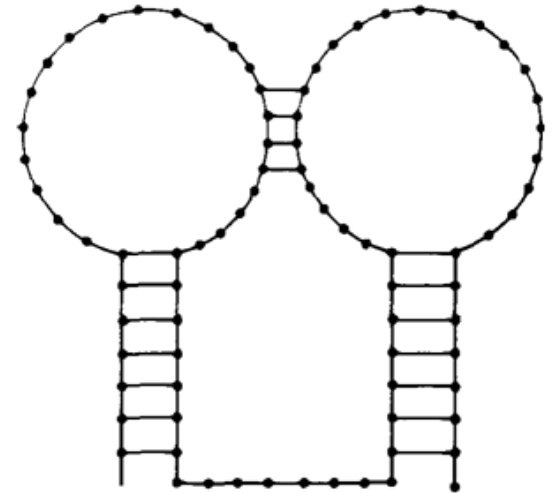
Pseudoknots

Pseudo-knot – not a knot

- why the name ?

Topologically like a knot

Would you expect them to occur ?



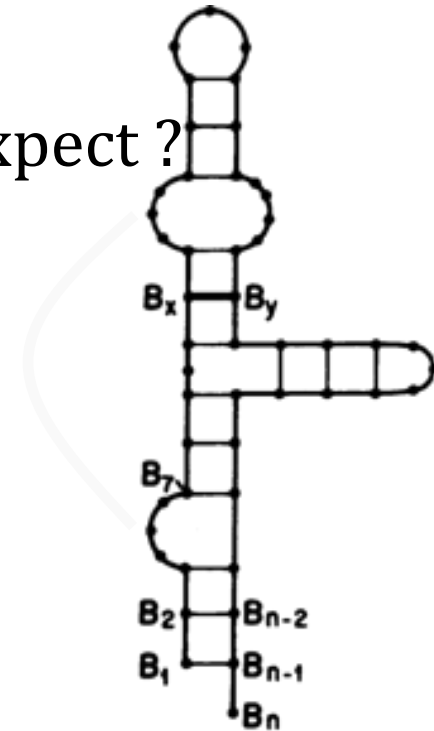
Pseudoknots

Given some unpaired bases, what would you expect ?

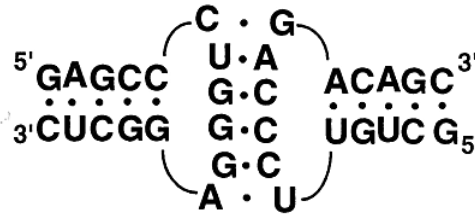
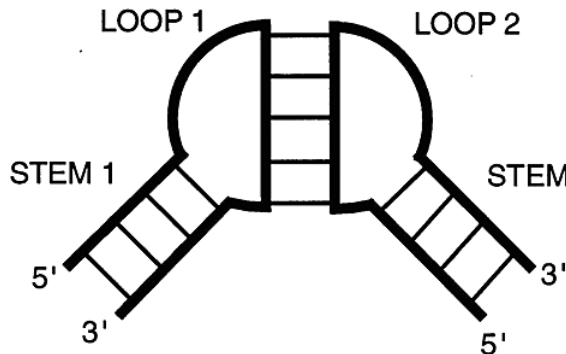
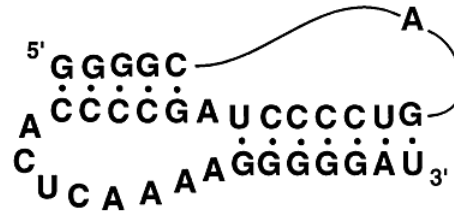
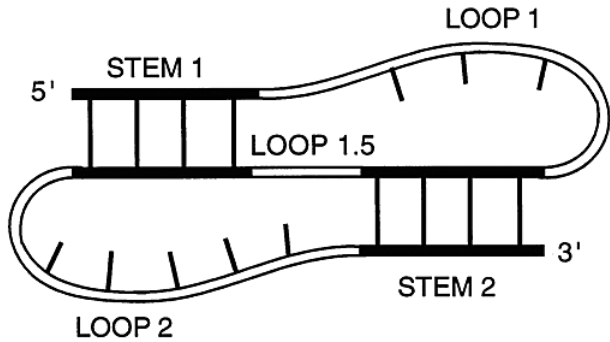
- solvate ?
- form more H-bonds ?
- pack bases against each other ?

Cannot (practically) be predicted

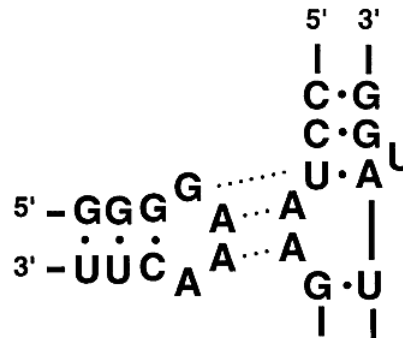
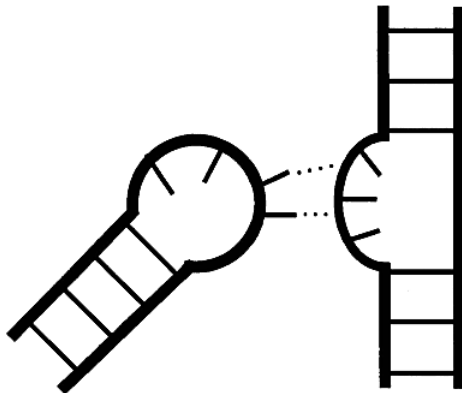
- order of steps in base-pairing methods



pseudoknots



kissing
hairpins



hairpin loop -
bulge

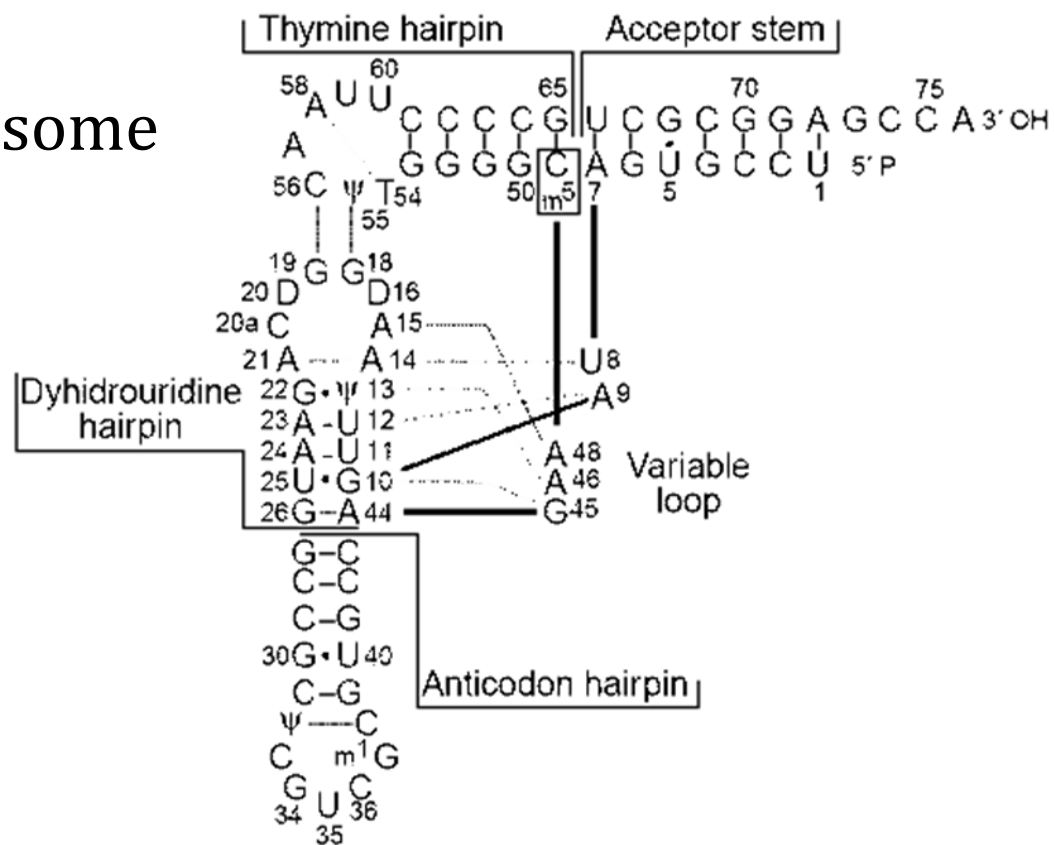
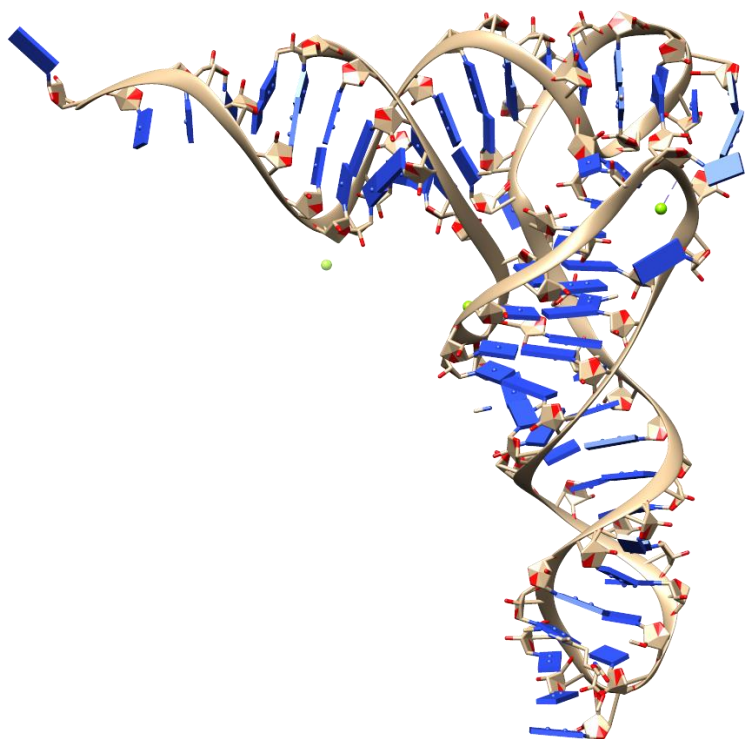
pseudoknots

Frequency of pseudoknots ?

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

Significant ?

- most structures will have some
- classic RNA example



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

- $3 \times$ GC versus $4 \times$ AU
- 9 H-bonds versus 8 H-bonds

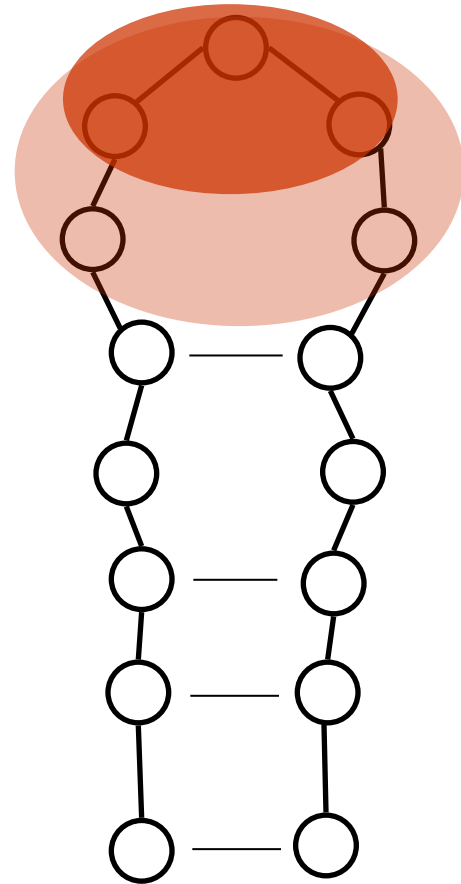
Scoring schemes – unpaired bases

Consider unpaired bases

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

Do these bases

- interact with each other ? solvent ?
- energy is definitely $\neq 0$



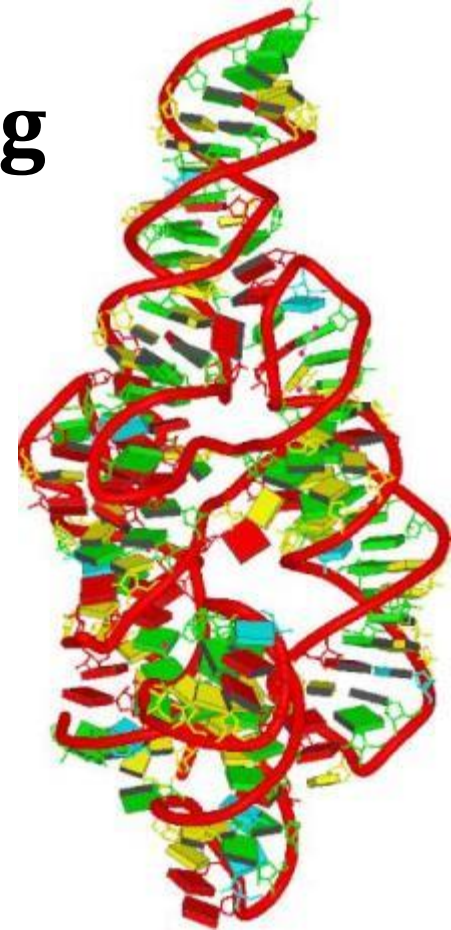
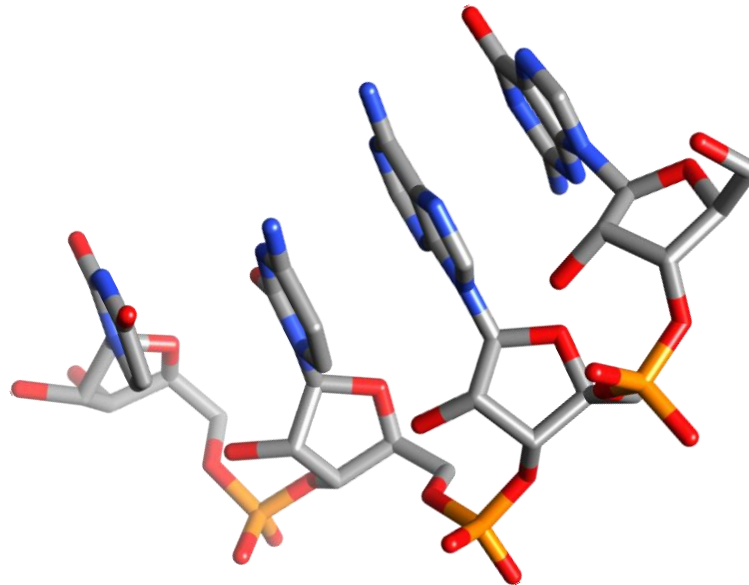
Scoring schemes - stacking

Bad assumption: each basepair is independent

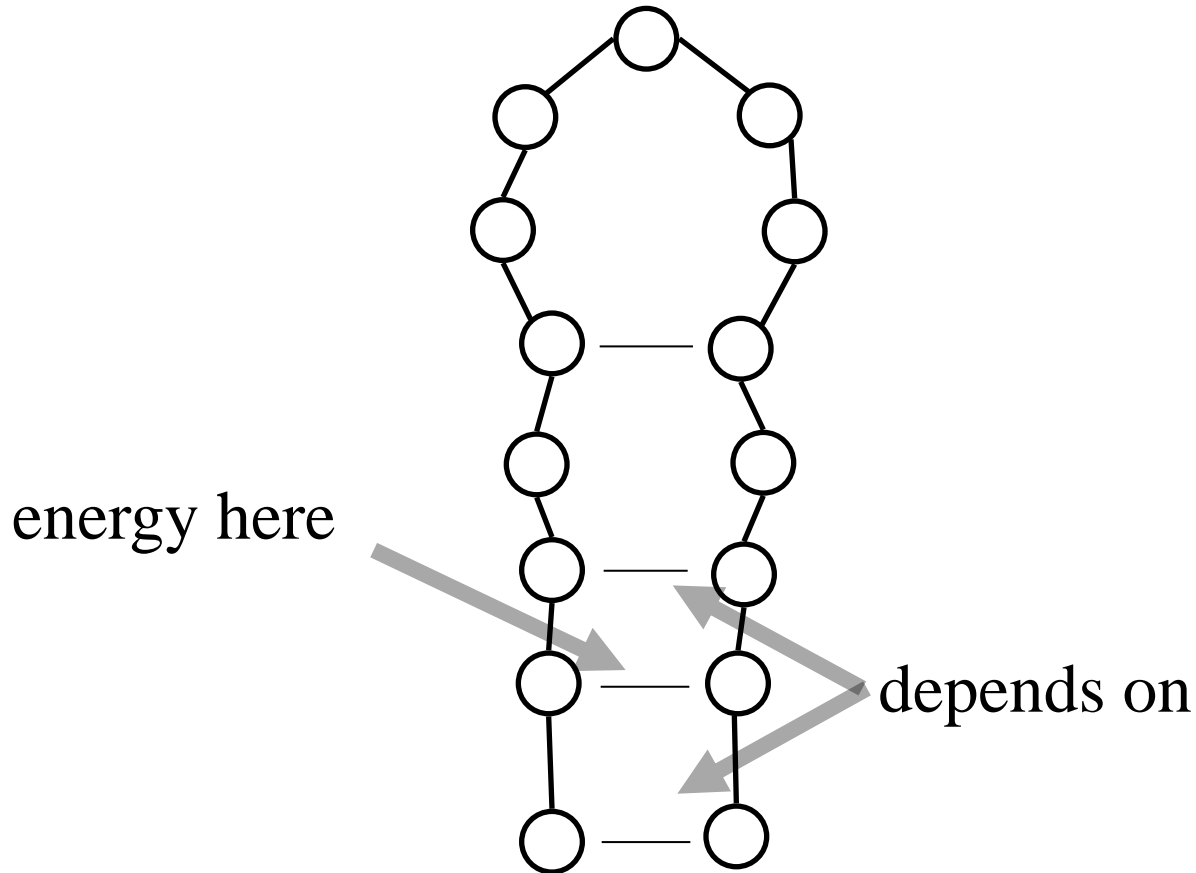
- $S(i,j) = \text{base-pair} + S(i+1, j - 1)$

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

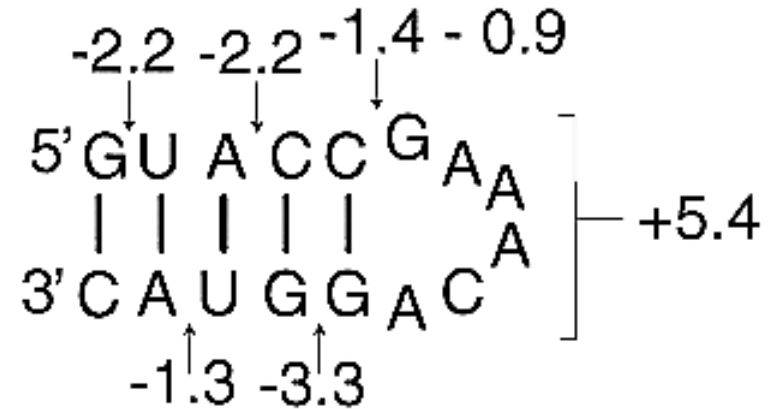
Previously we added

- GC + UA + AU + ...

Now

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

- terminal loop costs 5.4 kcal mol⁻¹



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 – alternative structures

Think of an "A"

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

Think of any base

- many possible good partners

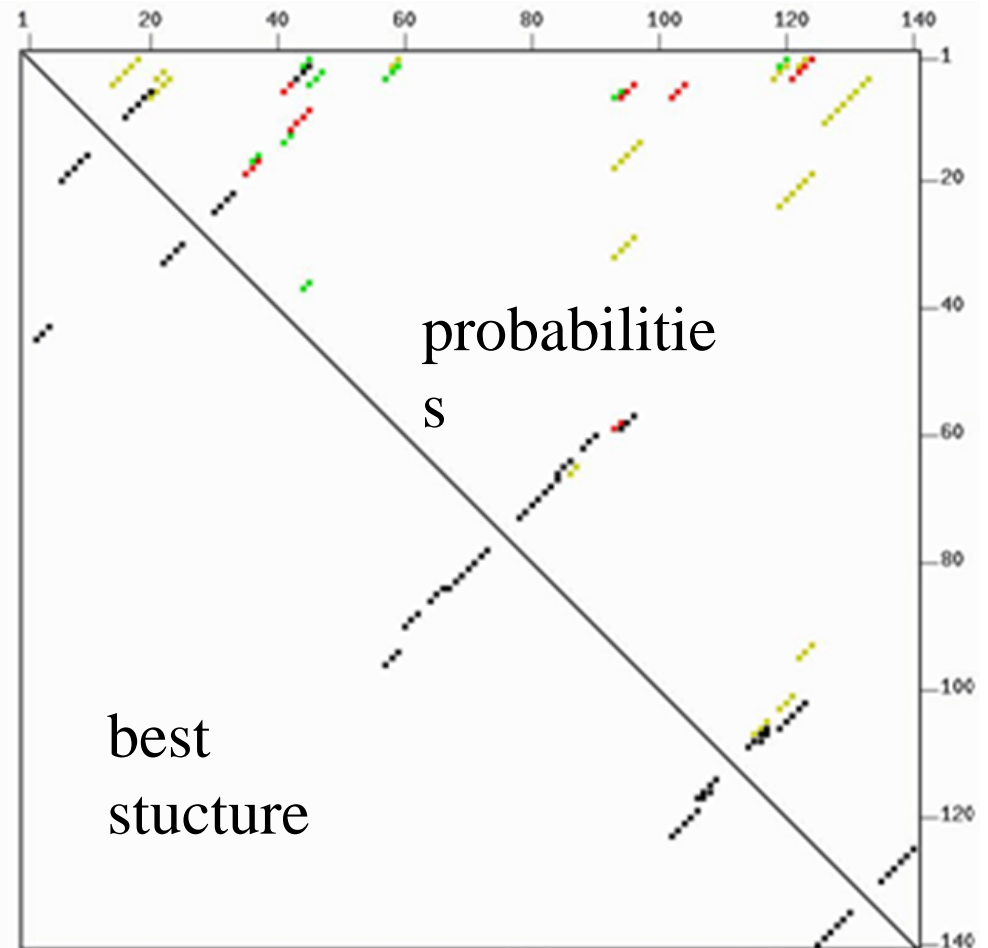
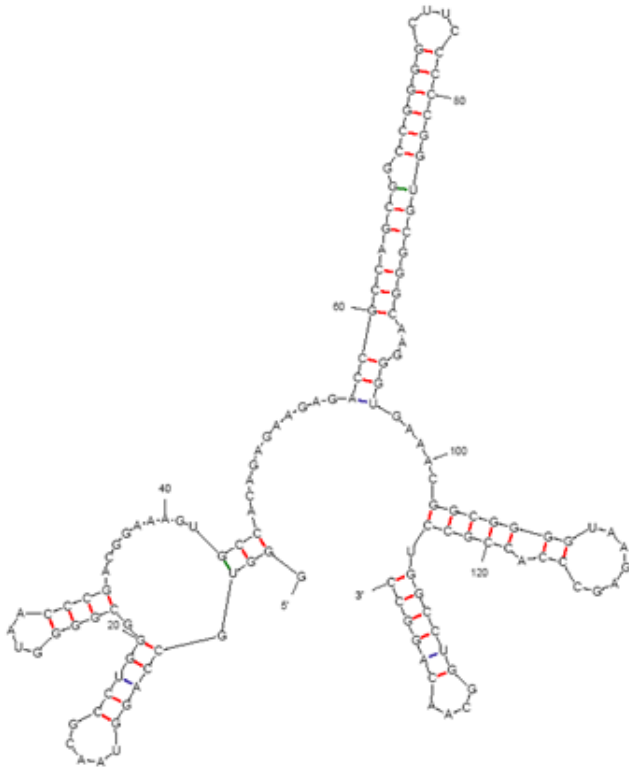
Consider whole sequence

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

Treat in terms of probabilities

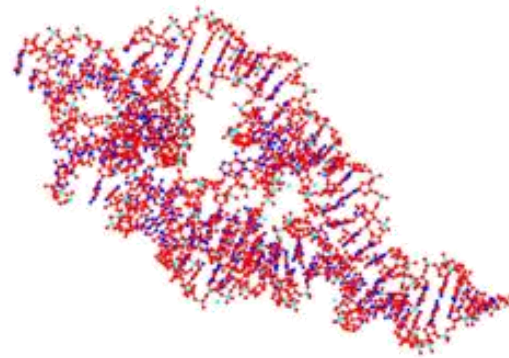
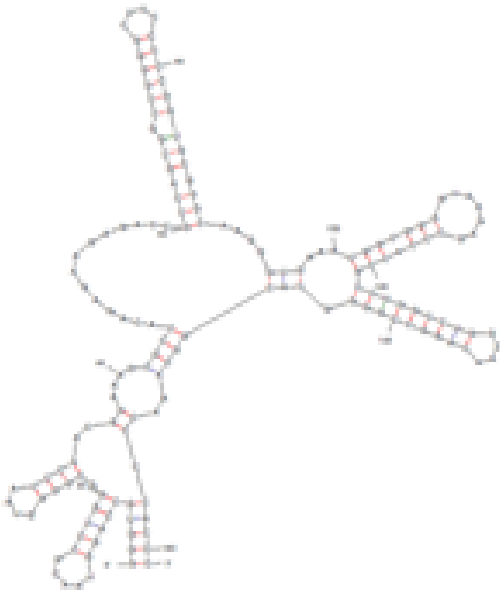
Probabilities

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

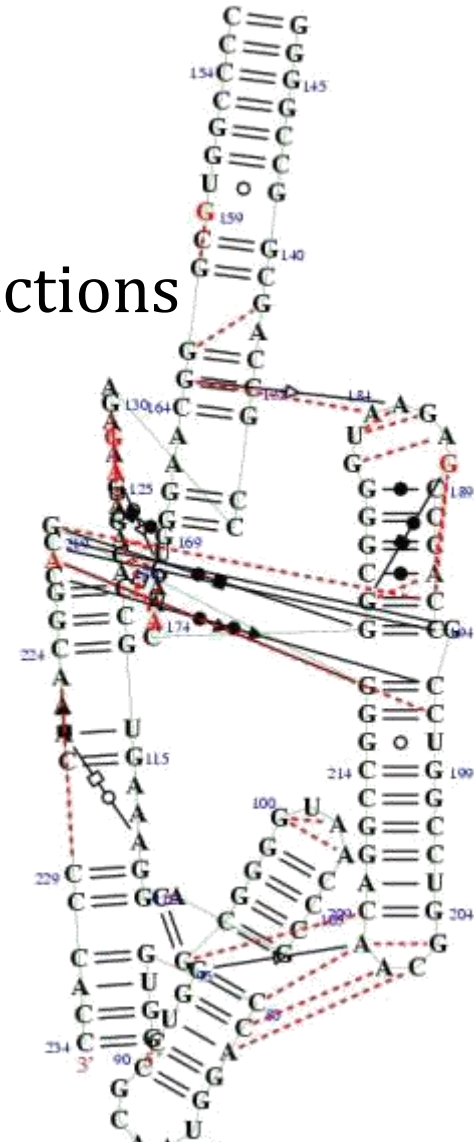


Reliability - Tertiary interactions

- miscellaneous H-bonds
- non-specific van der Waals
- Most larger RNA's have many tertiary interactions
- relatively compact



tertiary interactions
from crystal

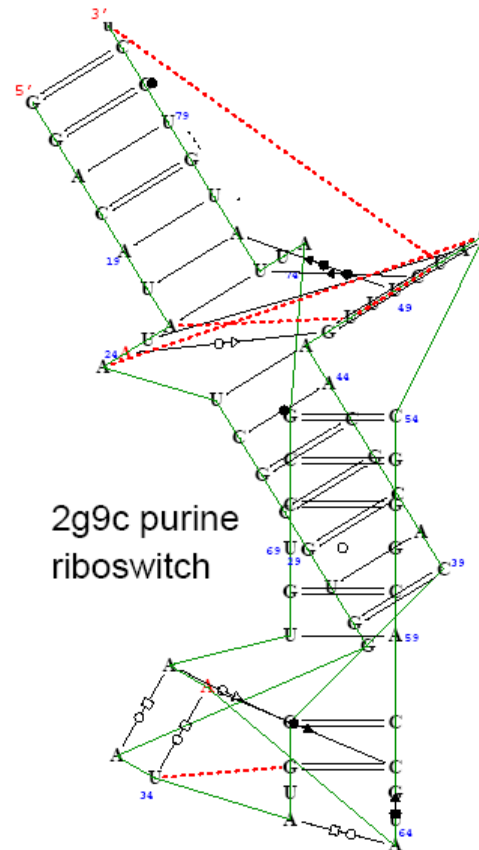


2D vs 3D

2g9c riboswitch



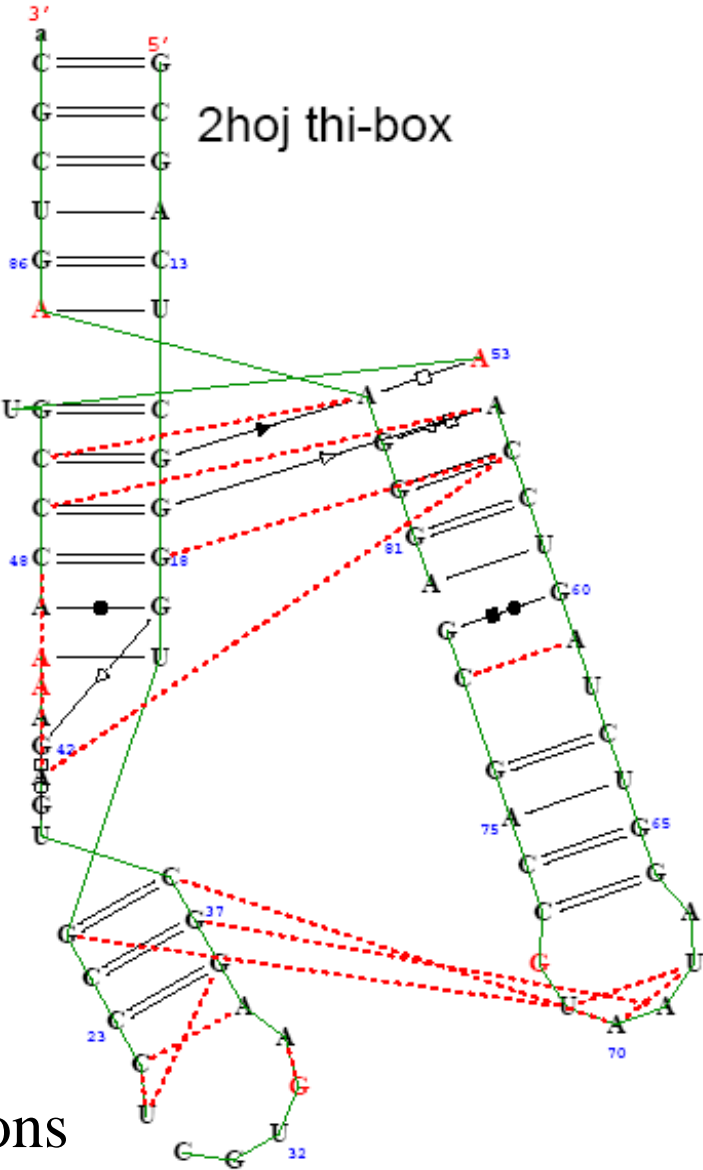
tertiary interactions
from crystal



2g9c purine
riboswitch

2D vs 3D

2hoj



tertiary interactions
from crystal

Reliability - summary

1. alternative structures with similar energies
 - if the second best guess is the correct one
 - you will not see it
2. tertiary interactions are not accounted for

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

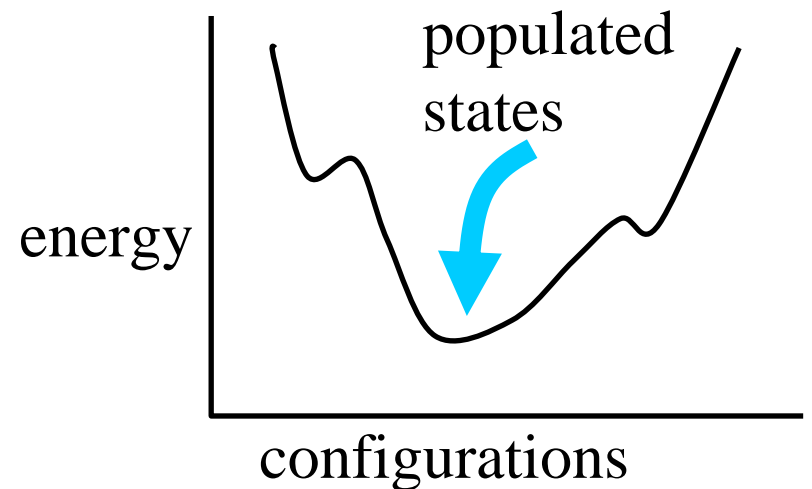
- kinetic trapping
- slow formation

Kinetic trapping

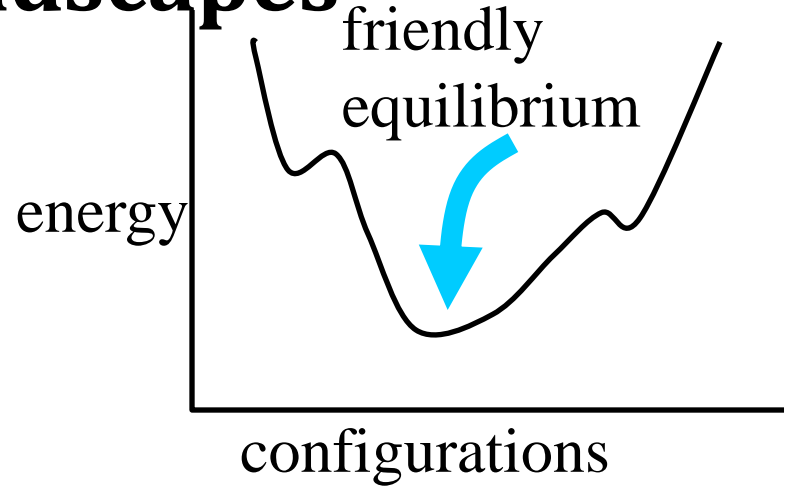
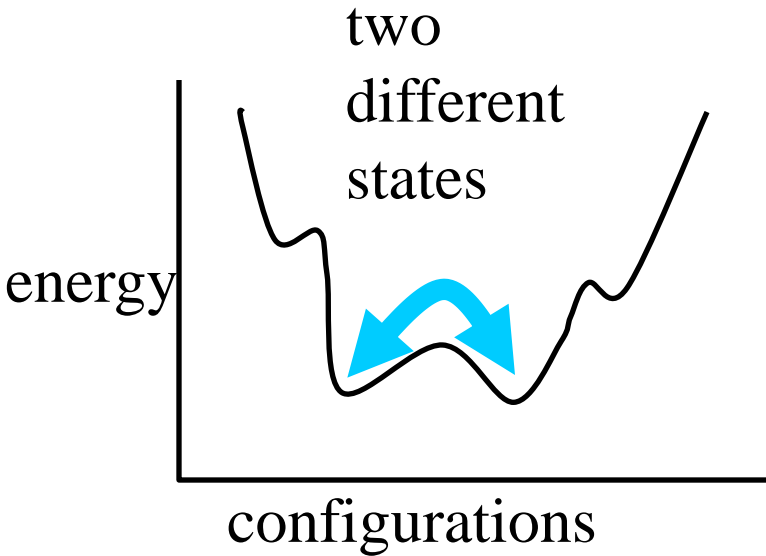
Term from protein world

Wherever the molecule is

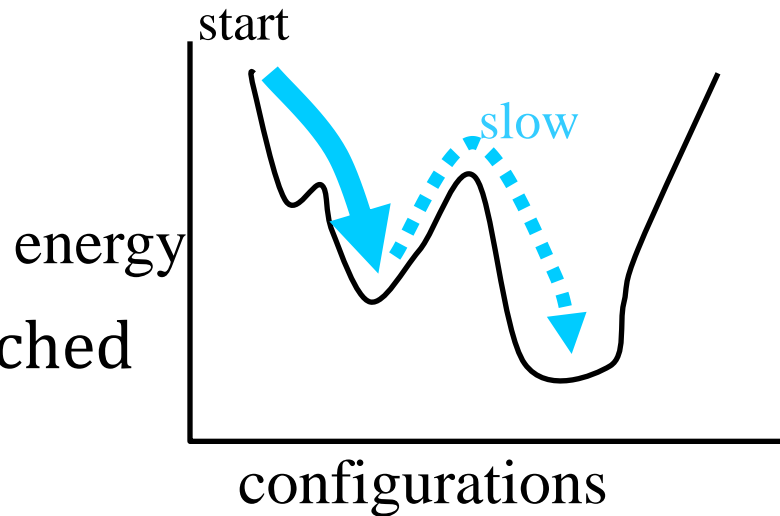
- it will probably go to energetic minimum
- less friendly landscape



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 ?

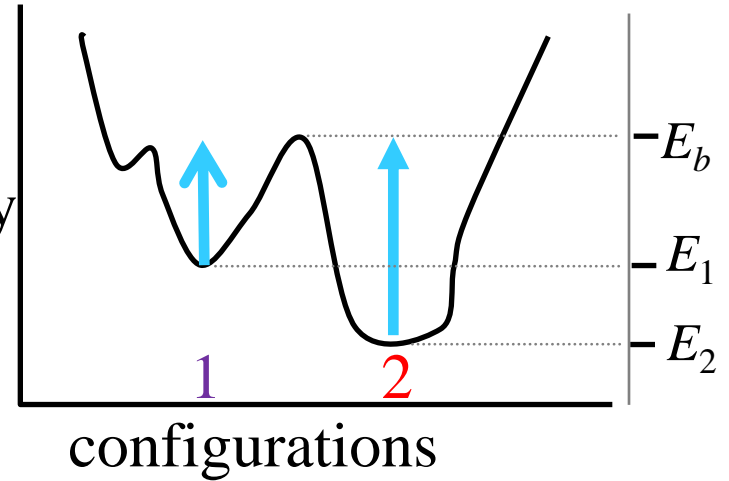
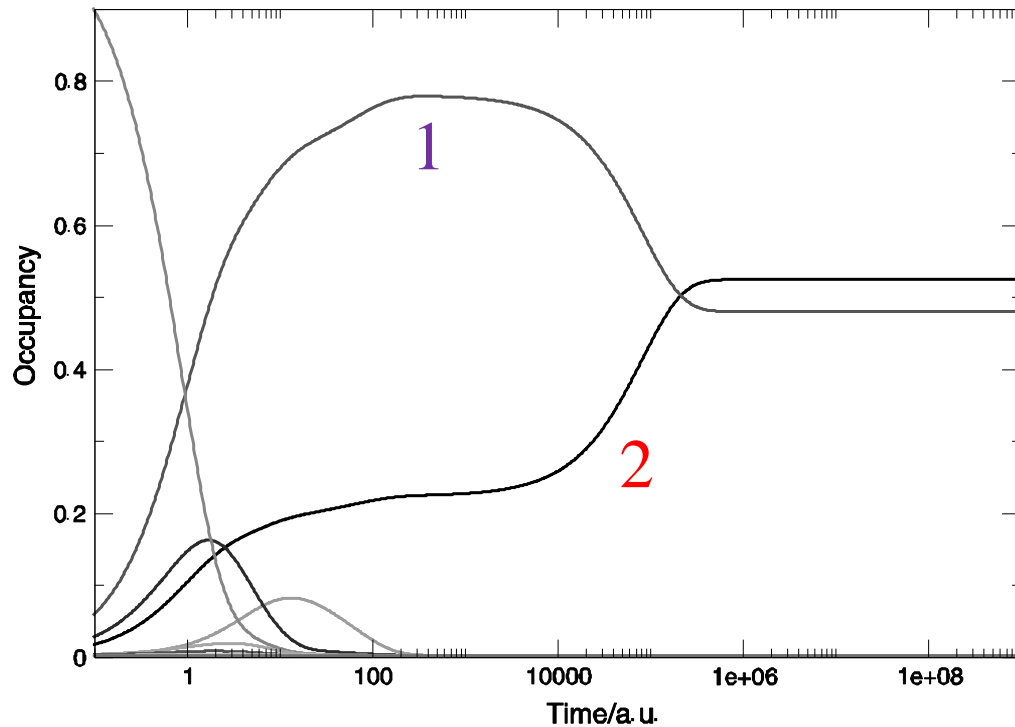
- not with methods so far

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

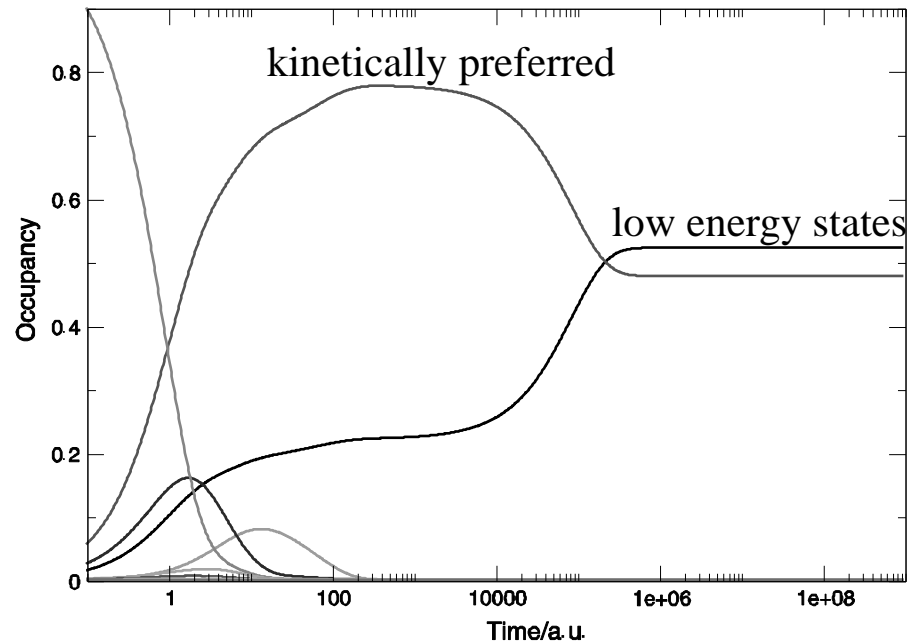


Implications

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)