

Introduction to Bioinformatics

April 29 2002

- Midterm Wednesday covers material through 26 April (BLAST)
- Review and beginning of RNA structure today



RNA Structure

RNA structure

- Representations
- Dotplots
- Folding

- Check out the RNA World
 - <http://www.imb-jena.de/RNA.html>



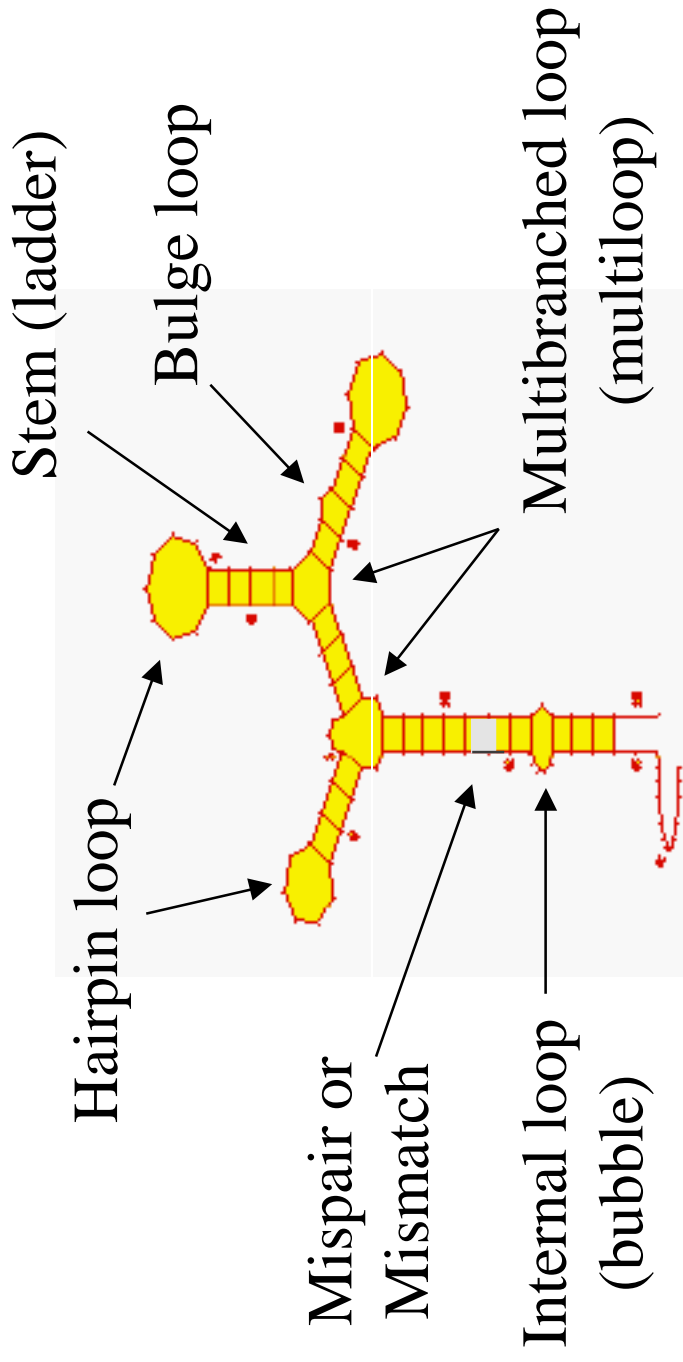
RNA Structure

RNA world hypothesis

- Before the “invention” of DNA and protein, early organisms relied on RNA for both genetic and enzymatic processes
- DNA was a selective advantage because it greatly enhanced the fidelity of genetic replication
- Proteins were a selective advantage because they make much more efficient enzymes
- Remnants of the RNA world remain today in catalytic RNAs in ribosomes, nucleases, polymerases, and splicing molecules
- See: Gilbert, W., “The RNA World”, Nature 319, 618, 1986.
<http://www.amsci.org/amsci/articles/95articles/cdeduvedu.html>
<http://www.panspermia.org/rnaworld.htm>

RNA Structure

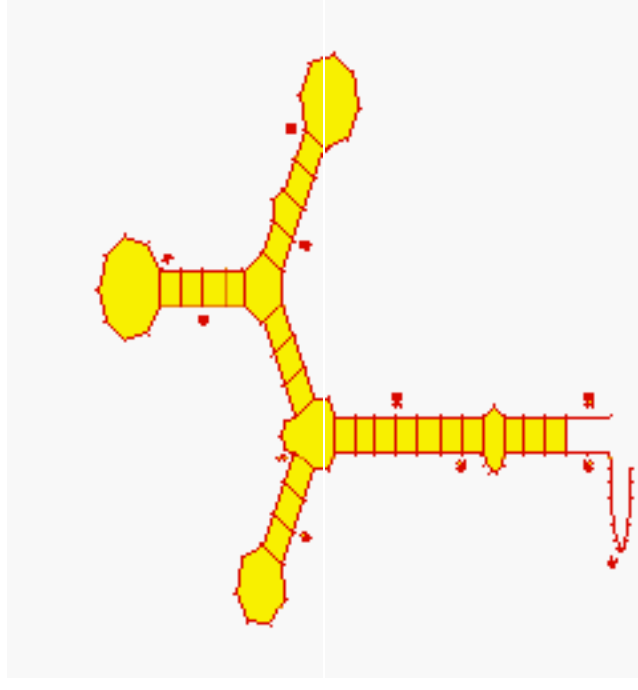
Terminology



RNA Structure

Representations - Stem/Loop ("Squiggles")

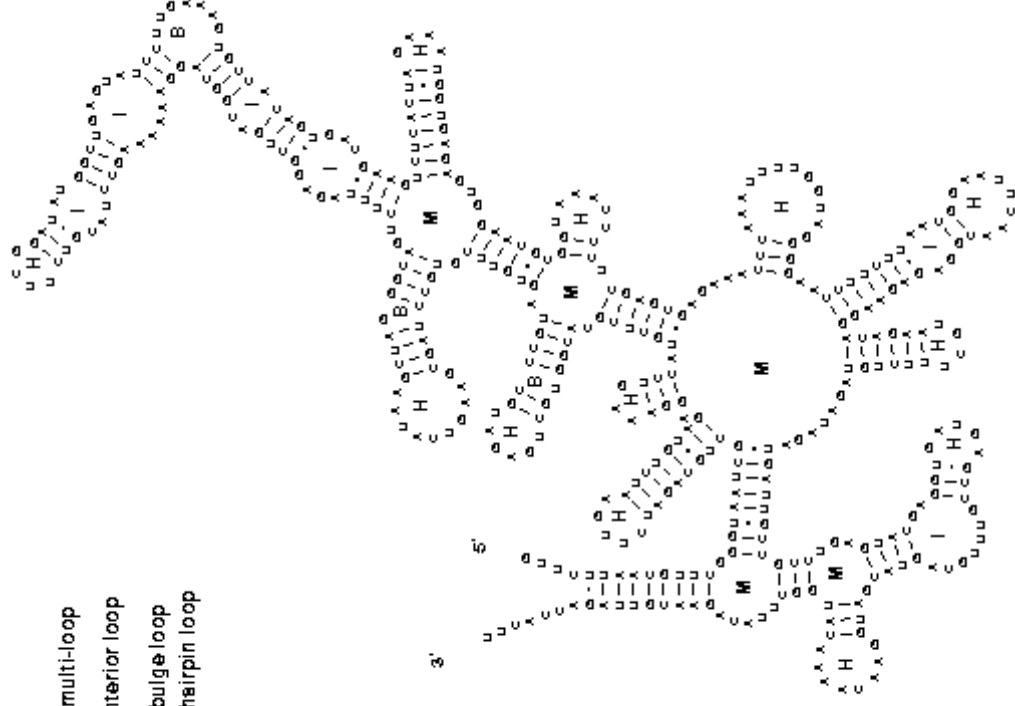
- Most common
- Biologically informative
- Difficult to compare



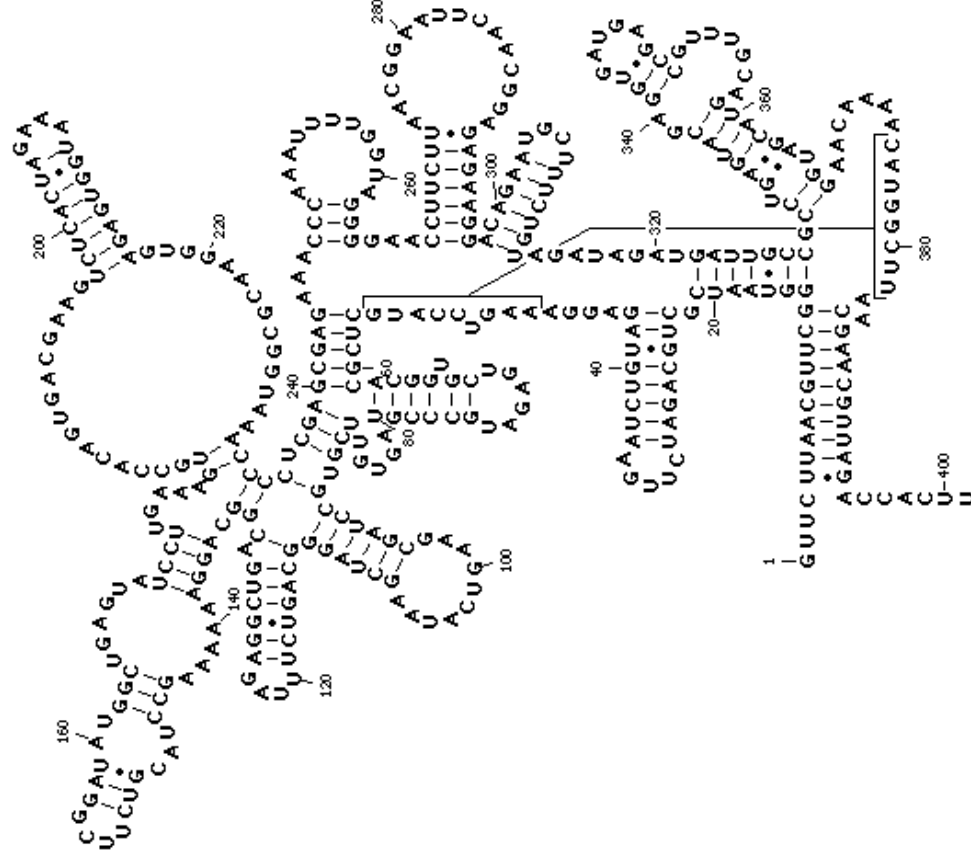
RNA Structure

Bacillus subtilis RNase P RNA

- M - multi-loop
- I - interior loop
- B - bulge loop
- H - hairpin loop



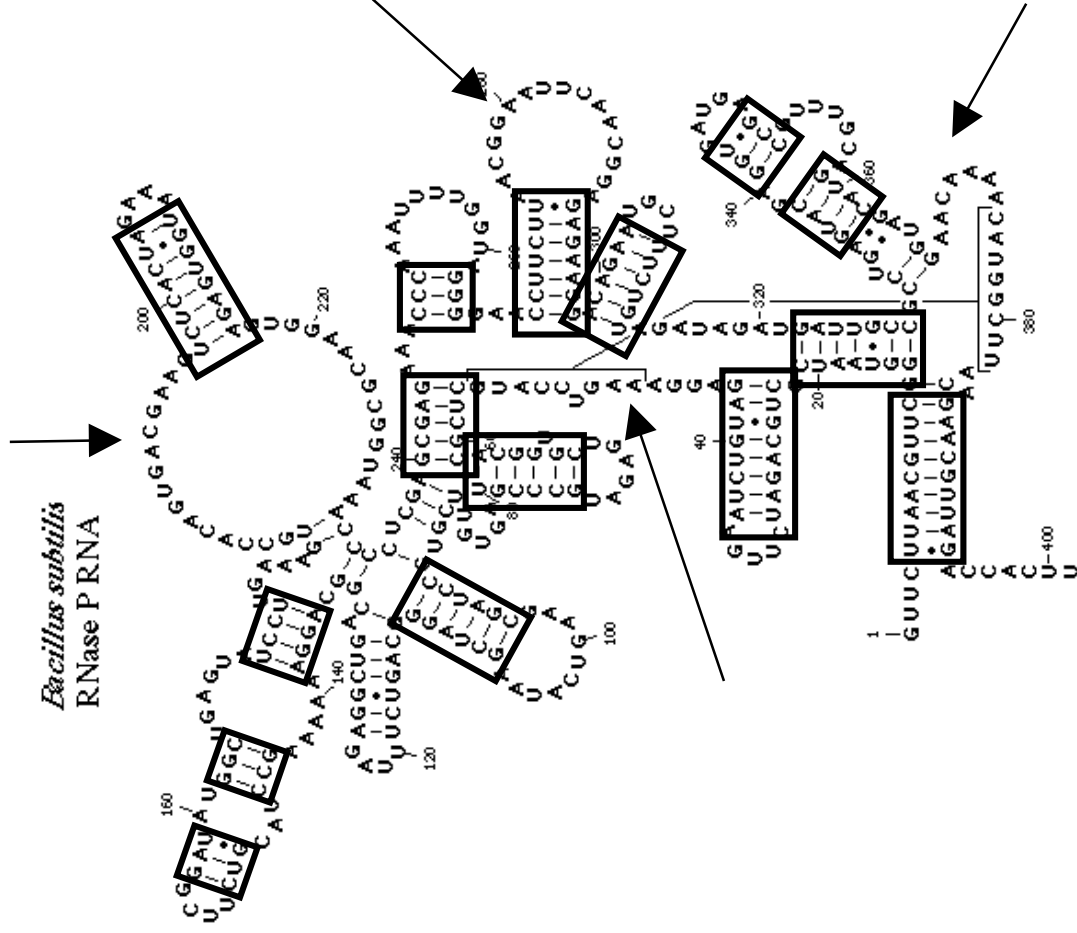
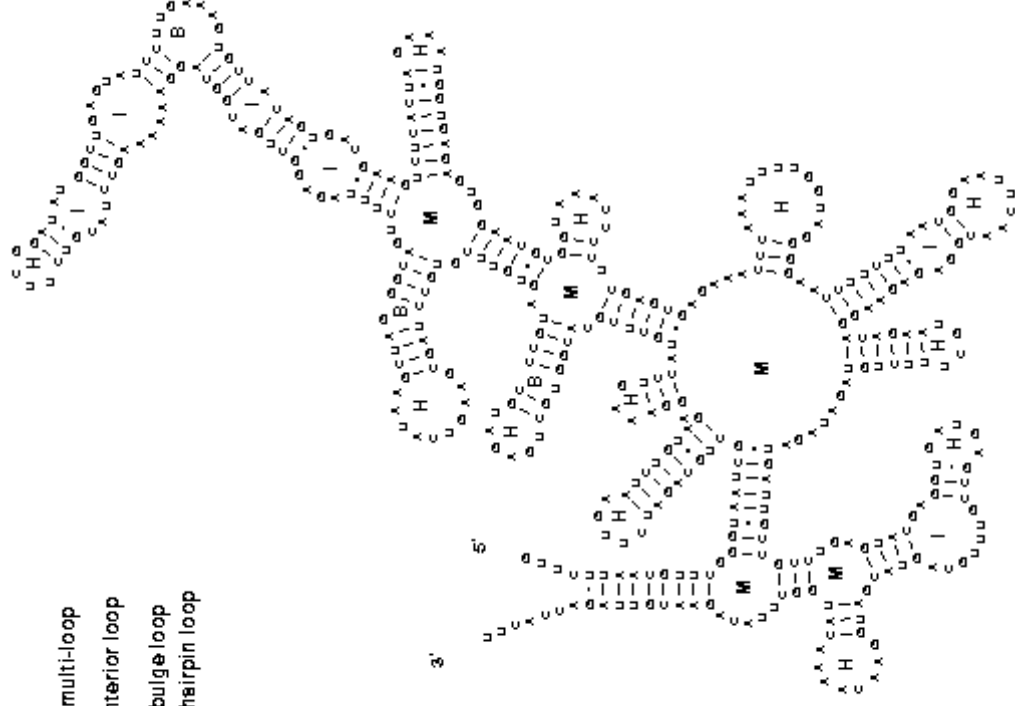
Bacillus subtilis
RNase P RNA



RNA Structure

Bacillus subtilis RNase P RNA

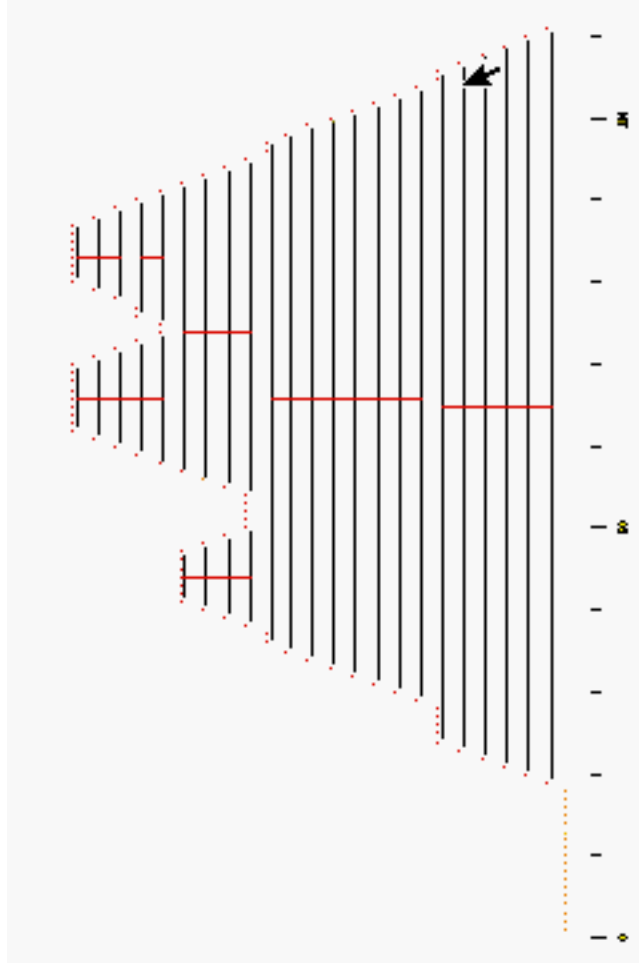
- M - multi-loop
- I - interior loop
- B - bulge loop
- H - hairpin loop



RNA Structure

Representations – Mountains

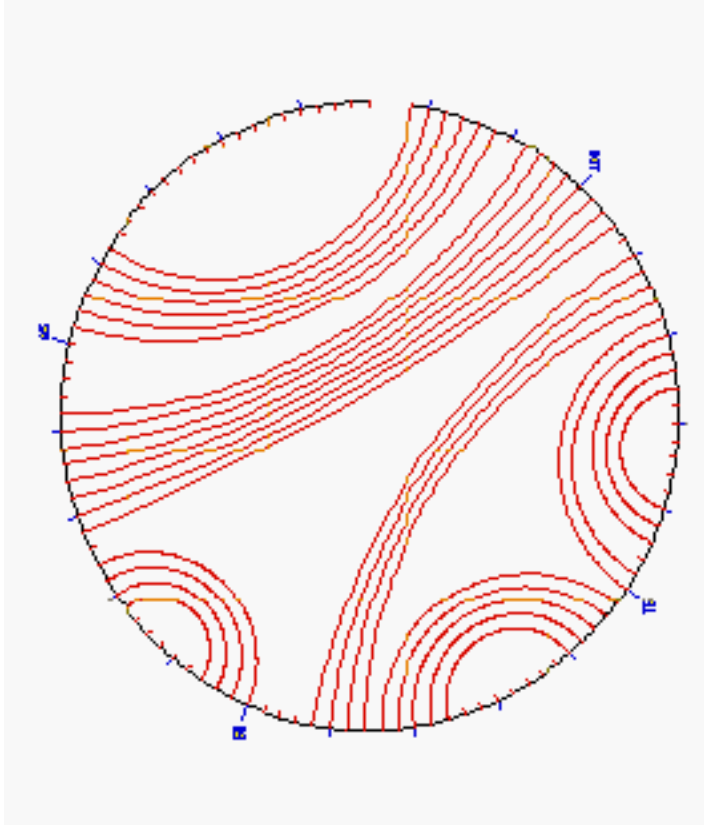
- Less common
- Is used in RNA literature
- Much easier to see similarity than “squiggles”
- Good for revealing pattern of nested stems



RNA Structure

Representations - Circles

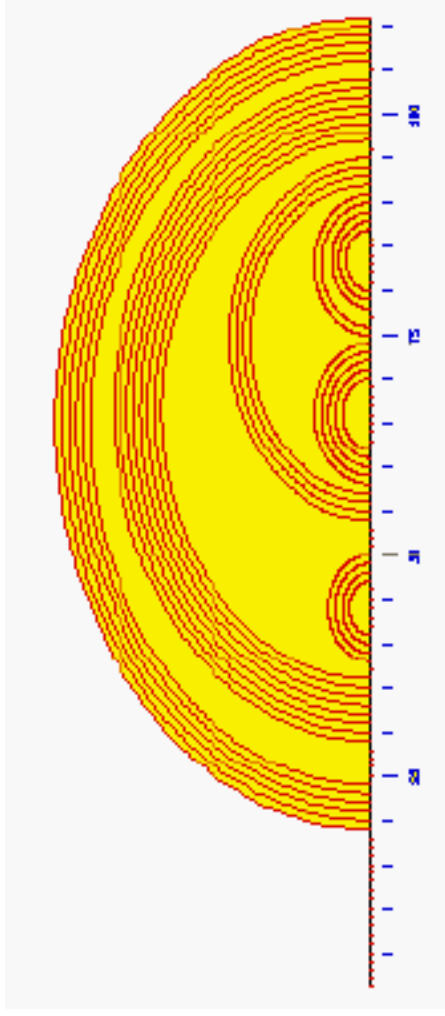
- Rarely seen
- Changes with length of RNA



RNA Structure

Representations - Domes

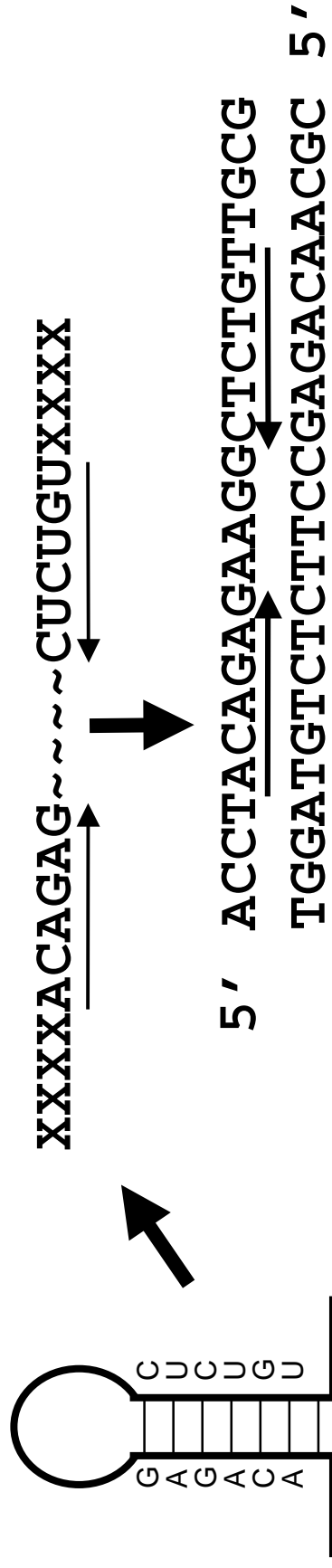
- Rarely seen
- Good for revealing pattern of nested loops



RNA Structure

Dotplot

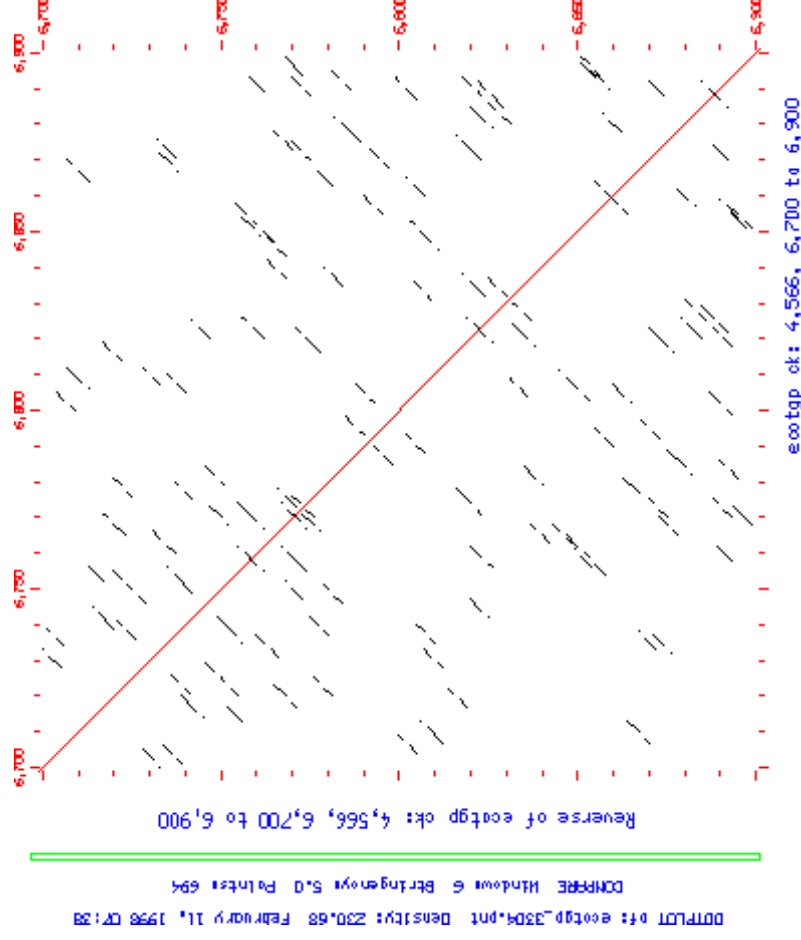
- Paired regions can be seen in dotplots
 - Use scores for complementary bases or plot vs reverse-complement sequence
- Use a windowing method to scan for stems
 - Window about the length of shortest stem
 - Can include crude energetics, GC=3 AU=2 GU=1



RNA Structure

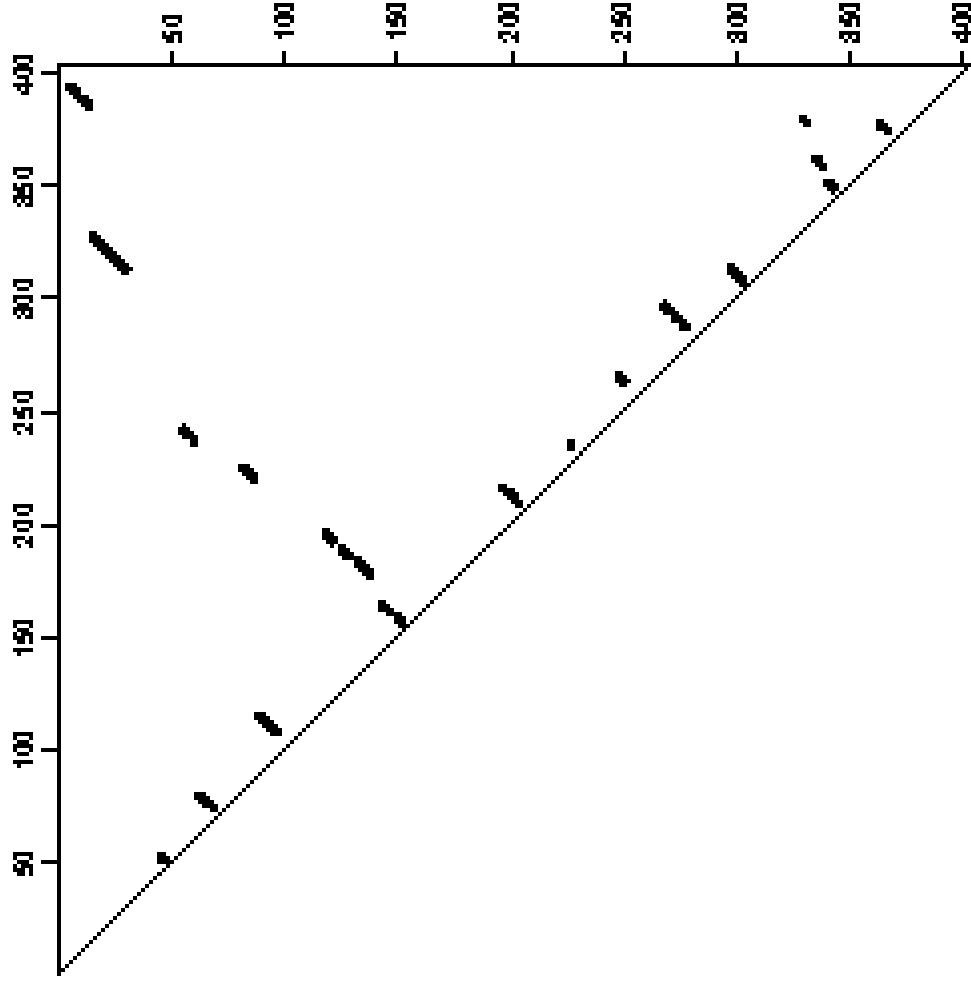
Dotplots

- Plot sequence vs. reverse complement
- Possible stems run perpendicular to axis of symmetry



RNA Structure

RNA Structure Dotplot - B. subtilis Rnase P



RNA Structure

Main Points

- RNA structure is dynamic in solution, i.e. constantly fluctuating between different folded states
- There are many alternative structures that are nearly identical in energy (both predicted and actual)
- Highly sensitive to solution conditions, e.g. salt and temperature
- Highly sensitive to protein binding
- Tertiary structure (e.g. pseudoknots are important)
- Biologically important structure may not have lowest predicted free energy, but it should be one of the lower ones - must look at sub-optimal structures

RNA Structure

Main Points

- **Three dimensional structure difficult to determine due to flexibility of molecule**
- **Most analysis of correctness must therefore rely on phylogenetically determined models**
- **Phylogenetic models look for invariant base pairs, but may not identify all unique structures**
- **Structural information also comes from nuclease digestion studies and sometimes crosslinking**

RNA Structure

Naive approach

- **GCG STEMLoop** program does a naïve search for stems
- **STEMLoop**
 - Calculates score over a window
 - Finds stems over a threshold score
 - Minimum/maximum loopsize
 - Sort by position or score
 - Can make dotplots

RNA Structure

STEMLOOP

Minimum Stem: 6 Minimum bonds/stem: 8.00 Maximum loop size: 20
Stems found: 72 Stems shown: 25
Average Match: 1.80 Average Mismatch: 0.00 Nibbling Threshold: 0.50

```
6850 CCGCCCCGGATCAGGTA AAAAG CAG 21, 30.0
    | | | | | | | | | | | | | |
6897 GCGGGCGTCGAGGACGTAGTT AGC 6
6823 CACCTCCATTGCAGGGATTGG TATT 22, 29.0
    | | | | | | | | | | | | | | | T
6875 GCGACGAAAATGGACTAGGCC CGCC 9
6841 TTGGTATTTCCGCCCCCGGATC AGGTAA 21, 28.0
    | | | | | | | | | | | | | | | A
6895 GGGCGTCGAGGACGTAGTTAG CGACGA 13
6823 CACCTCCATTGCAGGGAT TT 18, 26.0
    | | | | | | | | | | | | | | | G
6863 GACTAGGCCCCCGCCTTTA TG 5
6829 CATTGCAGGGATTGGTAT TTCC 19, 25.0
    | | | | | | | | | | | | | | | G
6875 GCGACGAAAATGGACTAGG CCCC 9
```

RNA Structure

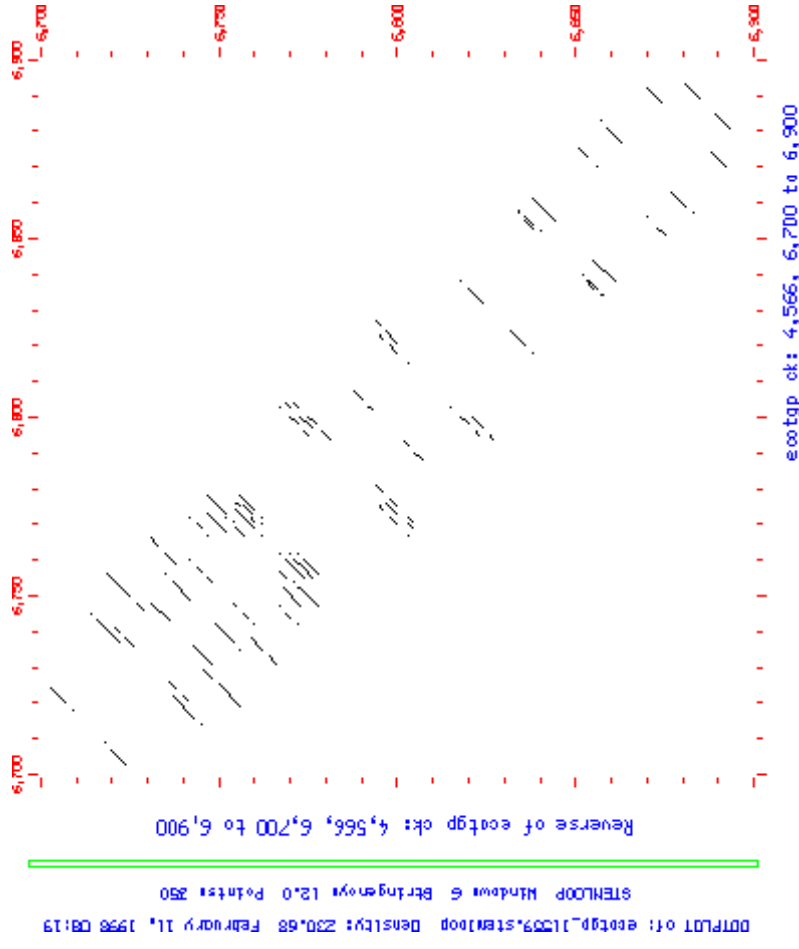
Naive approach

- **Problems with the naive approach**
 - Energetics are very crude
 - No bulges or bubbles
 - Complex optimization problem
- **Advantages**
 - Shows all stems not just lowest energy

RNA Structure

STEMLOOP

- Dotplot of structures



RNA Structure

Main Points

- RNA structure is dynamic in solution, i.e. constantly fluctuating between different folded states
- There are many alternative structures that are nearly identical in energy (both predicted and actual)
- Highly sensitive to solution conditions, e.g. salt and temperature
- Highly sensitive to protein binding
- Tertiary structure (e.g. pseudoknots are important)
- Biologically important structure may not have lowest predicted free energy, but it should be one of the lower ones - must look at sub-optimal structures

RNA Structure

Main Points

- **Three dimensional structure difficult to determine due to flexibility of molecule**
- **Most analysis of correctness must therefore rely on phylogenetically determined models**
- **Phylogenetic models look for invariant base pairs, but may not identify all unique structures**
- **Structural information also comes from nuclease digestion studies and sometimes crosslinking**

RNA Structure

Energetics

- The free energy of folded RNA molecules is a sum of several terms describing the base interactions in stem and loop structures

$$\Delta G = \Delta G_{stack} + \Delta G_{bulge} + \Delta G_{hairpin} + \Delta G_{internal} + \Delta G_{multibranch}$$

- ΔG_{stack} is the base-pairing/stacking energy for bases in stems, the others are all terms for loops
- Only ΔG_{stack} is favorable, all others destabilize secondary structures

RNA Structure

Energetics

- Base pair stacking (only favorable component)
- Depends on nearest neighbor

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

G:U base pairs are OK!

RNA Structure

Base-pair stacking

- Favorable energies come from base-pair stacking **NOT** from formation of base-pairs
- Un-paired bases make hydrogen bonds with water therefore there is no net change when they pair
- Favorable interactions come from electronic interactions between stacked bases
- Base-pair stacking is the **ONLY** favorable energy term in RNA folding

RNA Structure

Matches/Mismatches

- Get some favorable energy even if not hydrogen bonded due to stacking, for instance for a mismatch next to an A:U

5' AX 3'
3' UY 5'

	A	C	G	U
A	-0.8	-1.0	-1.7	-1.0
C	-0.7	-0.7	-0.7	-0.7
G	-1.5	-1.0	-1.0	-1.0
U	-0.8	-0.8	-0.8	-0.8

RNA Structure

Energetics

Loop Destabilization energy at 37 C

SIZE	INTERNAL	BULGE	HAIRPIN	SIZE	INTERNAL	BULGE	HAIRPIN
1		3.90		16	6.80	6.10	5.80
2	4.10	3.10		17	6.80	6.10	5.90
3	5.10	3.50	4.10	18	6.90	6.20	5.90
4	4.90	4.20	4.90	19	6.90	6.20	6.00
5	5.30	4.80	4.40	20	7.00	6.30	6.10
6	5.70	5.00	4.70	21	7.10	6.30	6.10
7	5.90	5.20	5.00	22	7.10	6.40	6.20
8	6.00	5.30	5.10	23	7.10	6.40	6.20
9	6.10	5.40	5.20	24	7.20	6.50	6.30
10	6.30	5.50	5.30	25	7.20	6.50	6.30
11	6.40	5.70	5.40	26	7.30	6.50	6.30
12	6.40	5.70	5.50	27	7.30	6.60	6.40
13	6.50	5.80	5.60	28	7.40	6.70	6.40
14	6.60	5.90	5.70	29	7.40	6.70	6.50
15	6.70	6.00	5.80	30	7.40	6.70	6.50

RNA Structure

Energetics

- **Tetraloops**
 - Exceptionally common 4 base long loops
 - >60% of loops in rRNA are
 - AUUU
 - CUUG
 - GAAA GAGA GCAA GCGA GGAA GUGA GUAA
 - UACG UCCG UUCG UUUU
 - Clearly more stable but exact energy unknown
 - Zuker gives -2 kcal/mol
- **No knots or pseudo-knots are allowed in secondary structure calculations**

RNA Structure

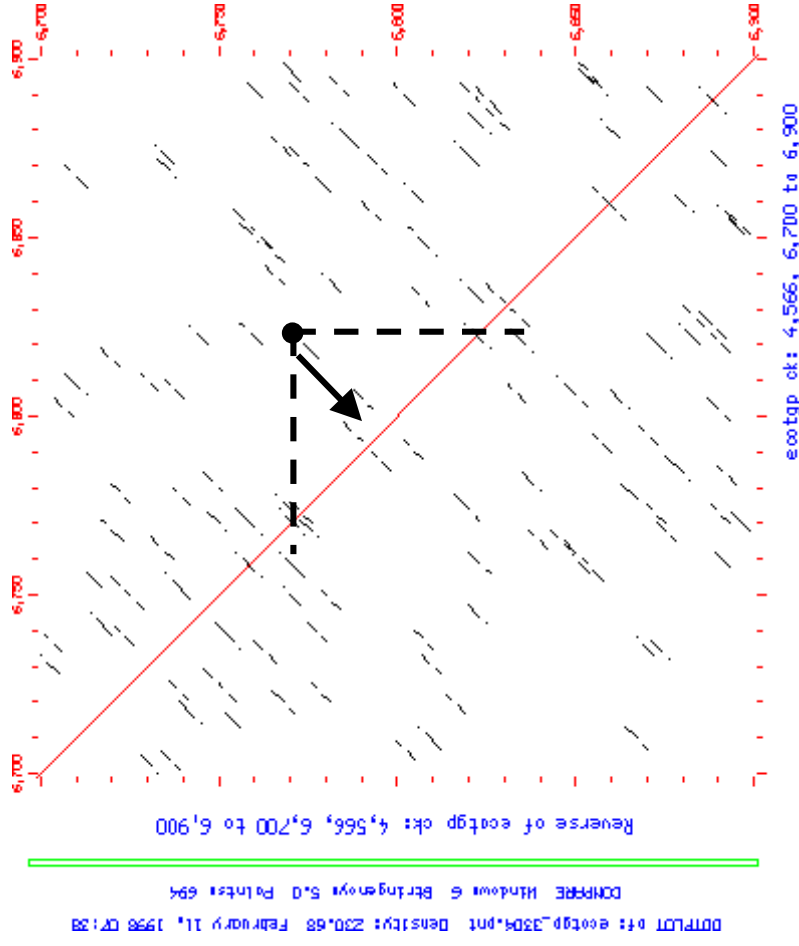
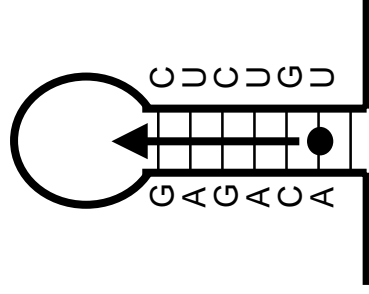
Algorithm

- **RNA folding is implicitly an N^4 algorithm**
 - N^2 dynamic programming to find the stems
 - N^2 dynamic programming to find the best combination
- **Zuker algorithm is N^3 due to approximations in searching for lopsided internal loops**
 - Note that very asymmetric internal loops will not be found with the default settings

RNA Structure

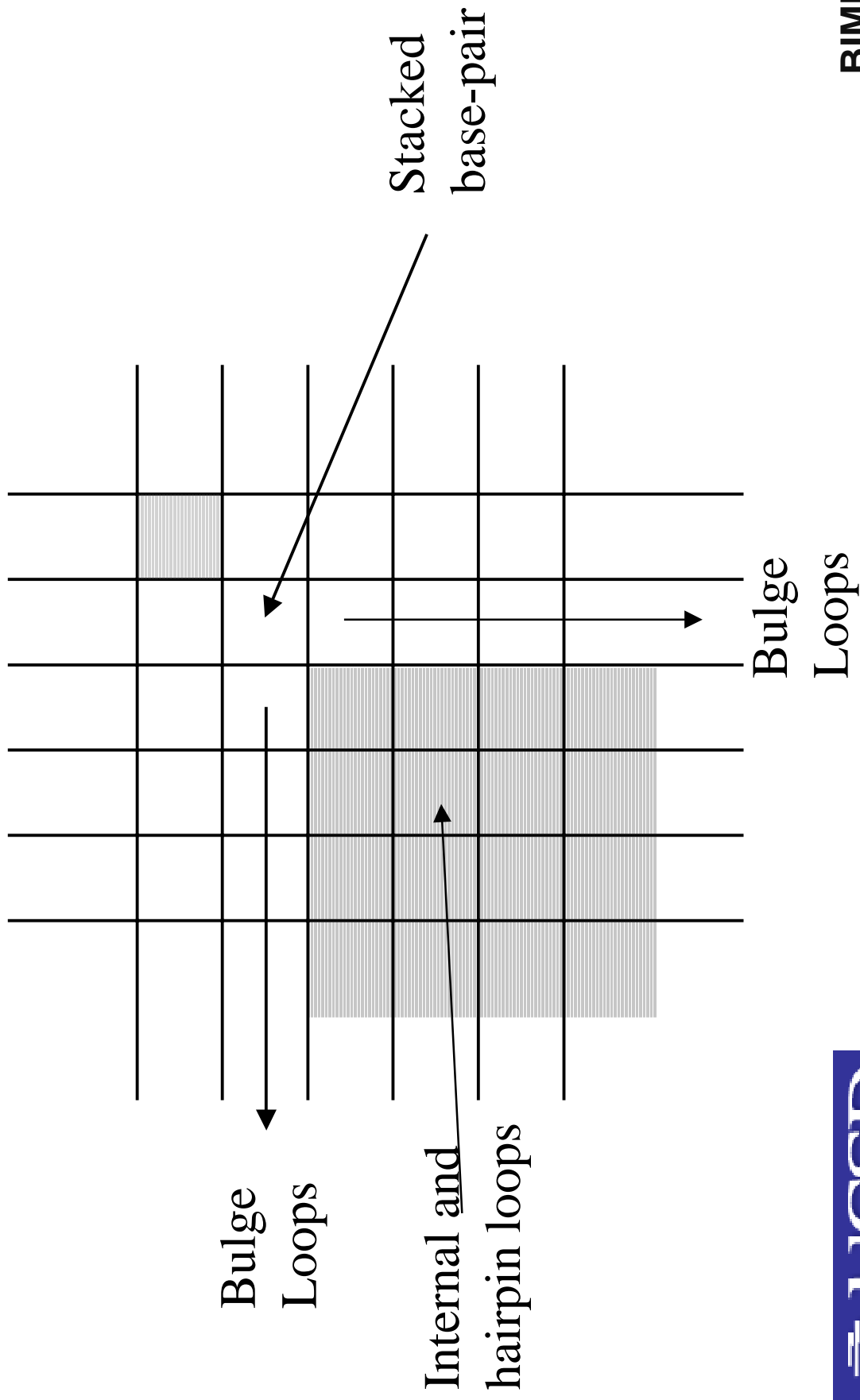
Algorithm

- For any stem, the stacked bases in the loop lay down and to the left



RNA Structure

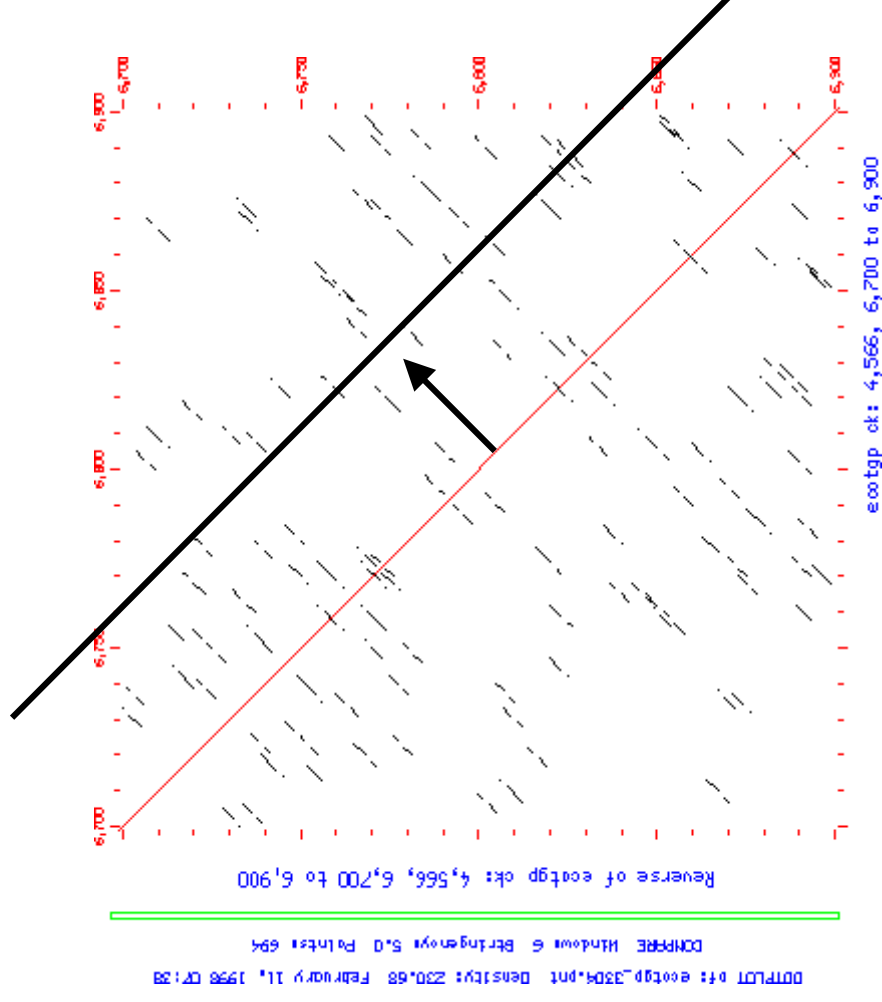
Algorithm



RNA Structure

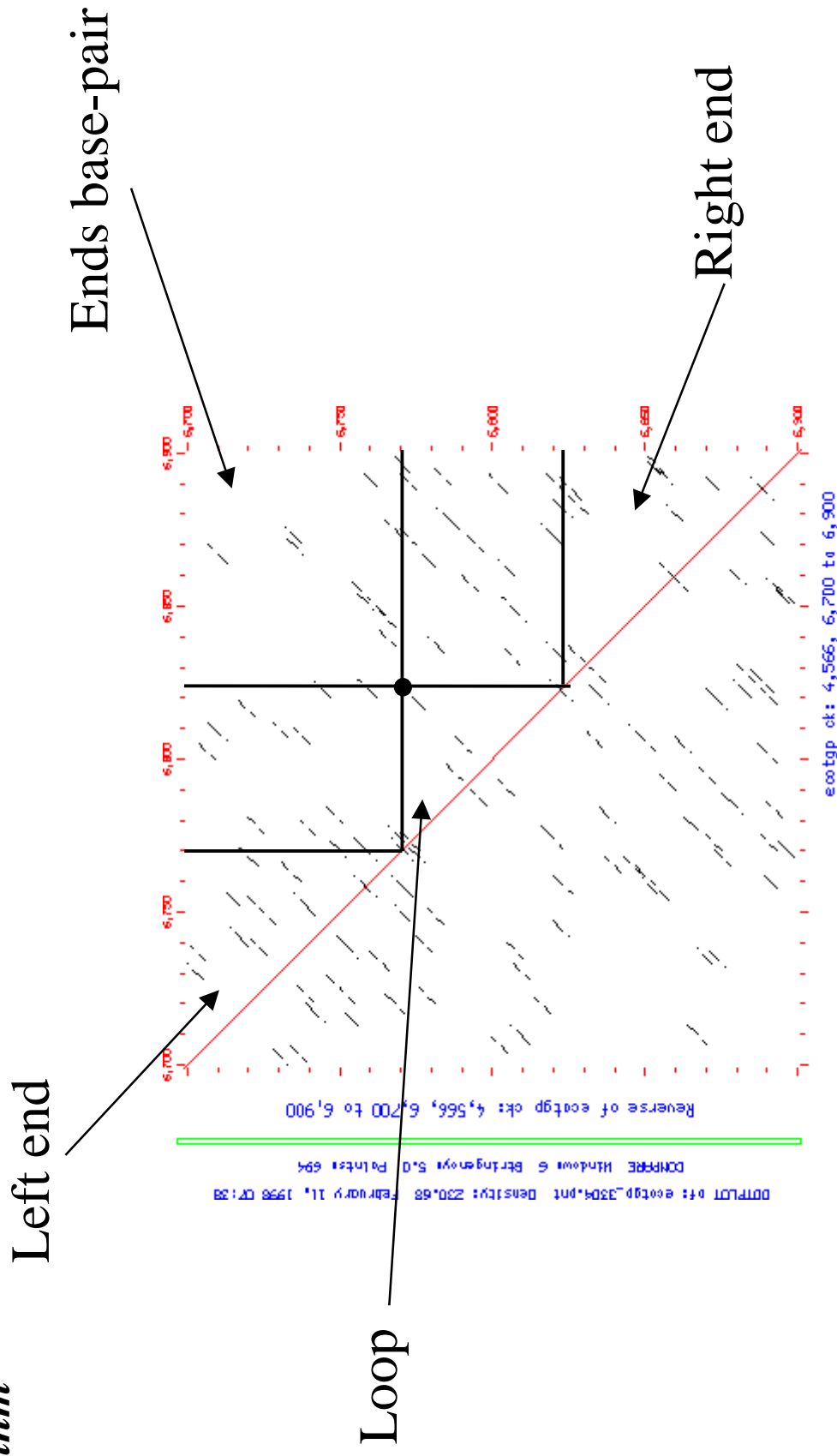
Zuker Algorithm

- Calculation proceeds from center towards edges
- Includes stacking, bulge, internal, and hairpin loop terms
- Start from center because the center line is location of hairpin loops



RNA Structure

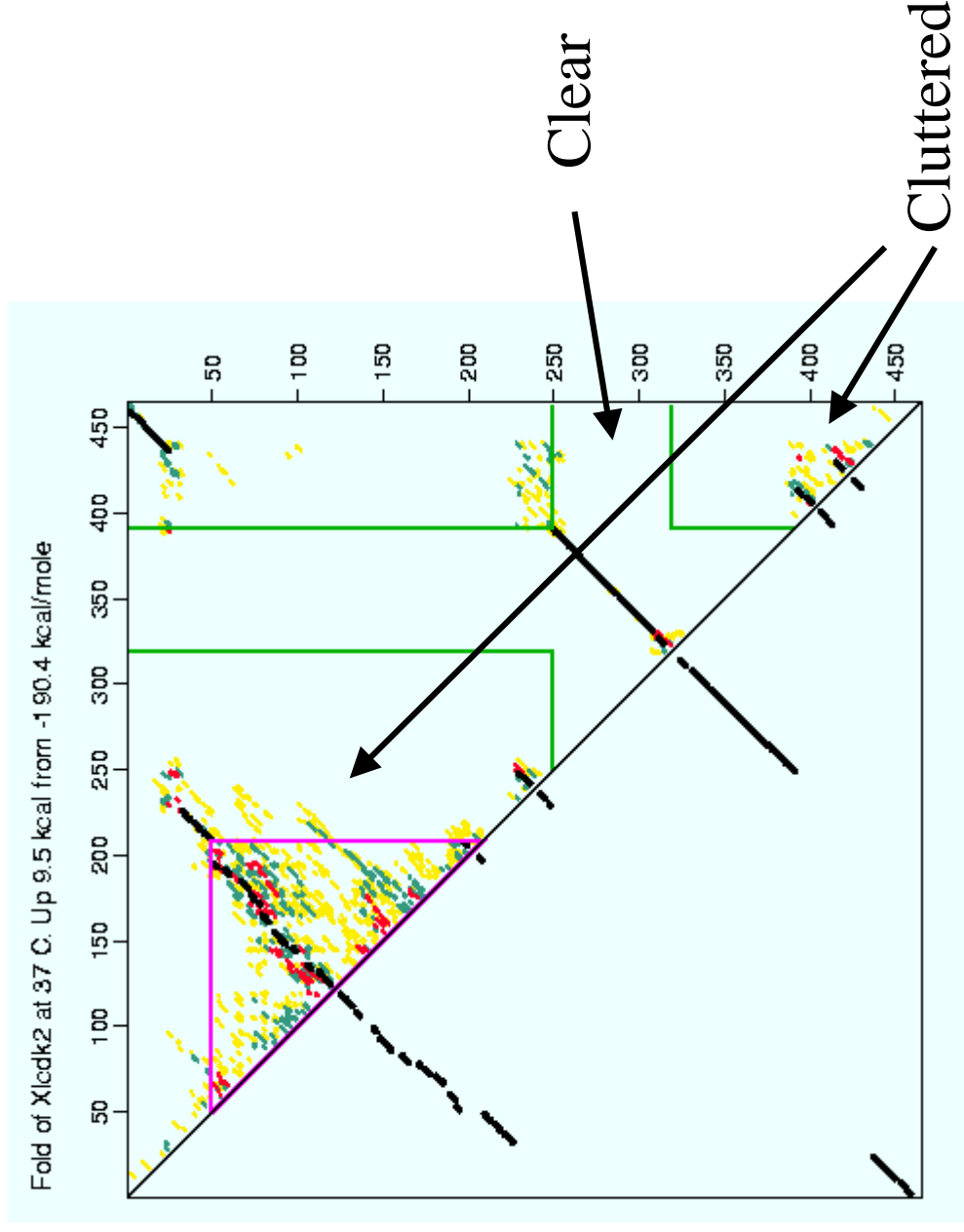
Algorithm



RNA Structure

Suboptimal structures

- Red 3.1 kcal/mol
- Blue 3.1 – 6.2 kcal/mol
- Yellow 6.2 – 9.3 kcal/mol



RNA Structure

Programs

- **MFOLD**
 - RNA secondary structure with suboptimal folding
 - Display results as mountains, domes, circles, squiggles Zuker's web site (includes server)
 - <http://bioinfo.math.rpi.edu/~zukerm/rna/>
 - Calculate energies for specified structure (efn server)
 - ssDNA structure prediction

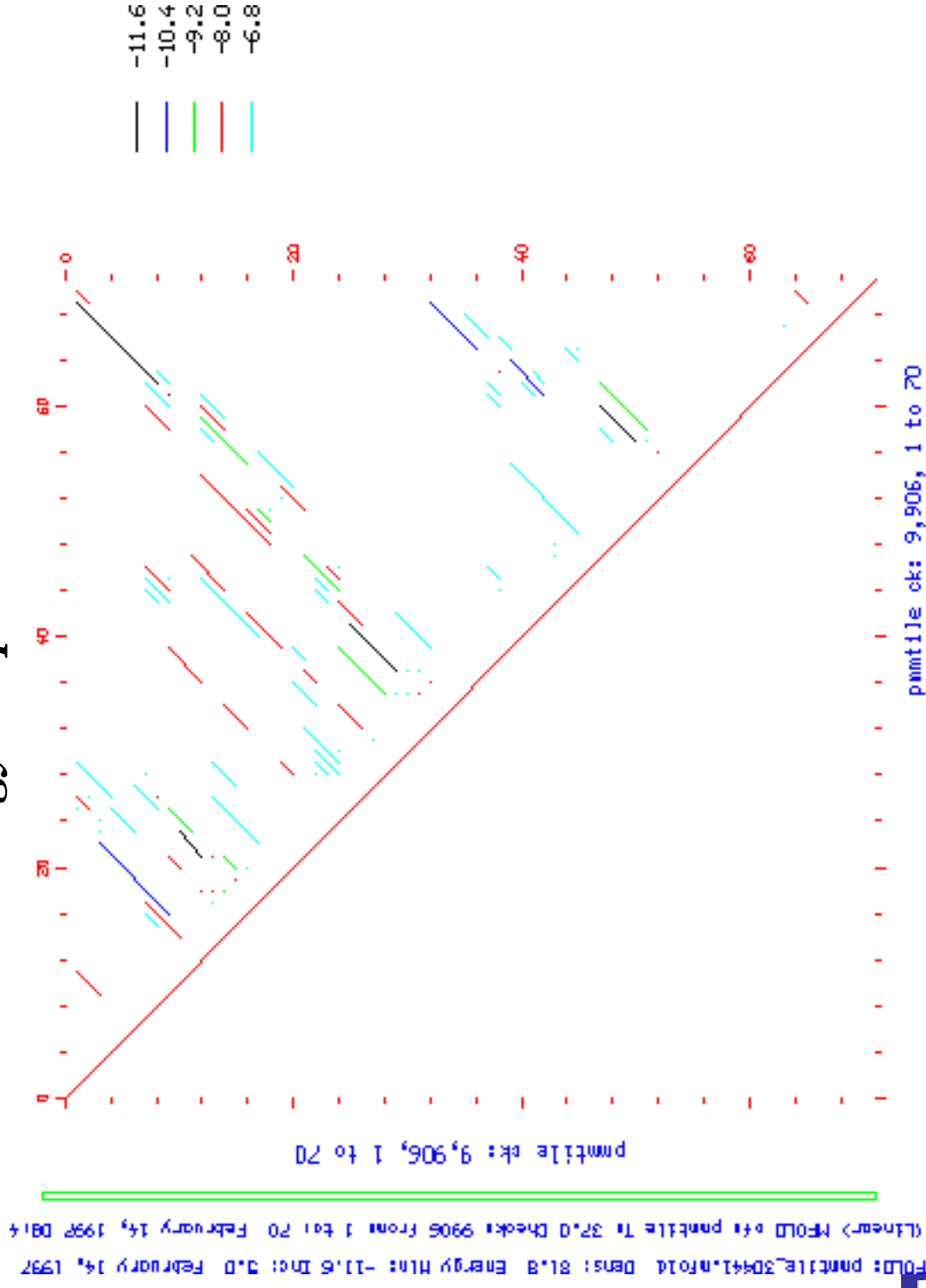
RNA Structure

MFOLD

- Uses forward-backward algorithm to get structures close to optimal parameters
- **Temp=x** Temperature for folding
- **LOPsidedness=30** maximum lopsidedness of an interior loop
- **FORCe=i,j,k** forces k consecutive base pairs, starting with base pair i, j
- **FORCe=i,0,k** forces k consecutive bases, beginning with i, to base pair
- **PREVent=i,j,k** prevents k consecutive base pairs, starting with base pair i, j
- **PREVent=i,0,k** prevents k consecutive bases, starting at i, from base pairing
- **CLOSeDexcise=i,j** excludes bases i+1 - j-1 from folding, forcing base pair i, j
- **OPENexcise=i,j** excludes bases i - j from folding, ligating base i-1 to j+1

RNA Structure

GCG PLOTFOLD - Energy Dotplot



RNA Structure

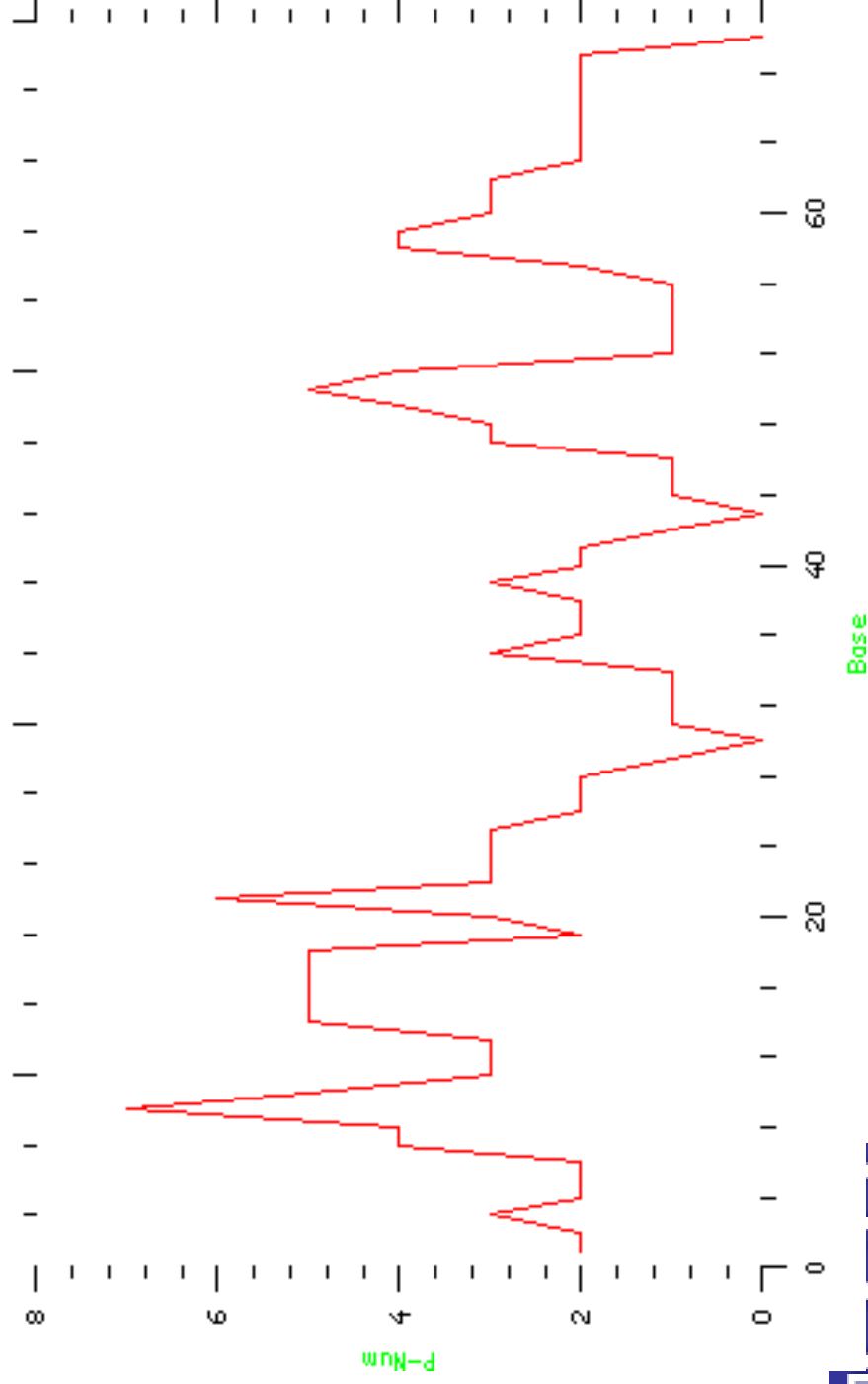
Evaluation

- **Biological RNAs (with important structure) are difficult to distinguish from random RNAs**
 - Same number and length of stems and loops
 - Same GC content of stems
 - Same free predicted free energy
- **Biologically important structures are exceptional in lacking competing structures**
 - this insures that the structure will be present regardless of the net DG
- **PNUM plot shows number of alternative structures within energy increment**
- **Agrees well with phylogenetic predictions, but most effective for large molecules**

RNA Structure

Evaluation - PNUM Plot

P-Num plot of: pmmtile_30441.mfold February 14, 1997 10:48
Min Energy: -11.6 Energy Inc: 3.0 Density: 60.87
<Linear> MFOLD of: pmmtile T: 37.0 Check: 9906 from 1 to: 70 February 14, 1997 08:45



RNA Structure

Phylogeny based prediction

- Phylogenetic structure depends on covariance analysis
- Look for bases that always mutate in a concerted way, e.g. A:T changing to C:G
- Only defines conserved portion of structure, not unique portions
- Many structures available for (see beginning of lecture)
 - Ribosomal RNA
 - Ribonuclease P RNA
 - RNPs
 - Introns

RNA Structure

Phylogeny based prediction

- **Mutual Information**

- Correlation between columns i and j in a multiple alignment is measured by the Mutual Information, H

$$H(i,j) = \sum f_{ij}(N_1, N_2) \log_2 f_{ij}(N_1, N_2) / f_i(N_1) f_j(N_2)$$

over all bases $N_1:N_2 \in \{A:T, C:G, G:C, T:A\}$

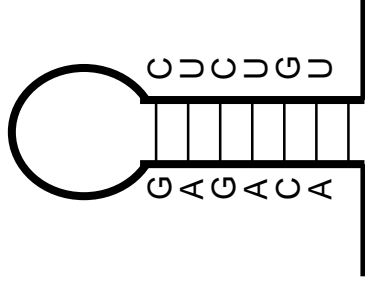
$f_{ij}(N_1, N_2)$ is the joint frequency of N_1 and N_2 in columns i and j

- 0 when positions are random, 2 when in perfect agreement with base-pairing

RNA Structure

Phylogeny based prediction

Consider two columns that are base paired, for instance the A:U pair at the base of the stem. Over time mutations will tend to maintain the base-pairing, even though the bases change. You might observe the following in a group of related sequences:



...	A	...	U	...
...	U	...	A	...
...	A	...	U	...
...	U	...	A	...
	<i>i</i>		<i>j</i>	

$$H(i,j) = \sum f_{ij}(N_1, N_2) \log_2 f_{ij}(N_1, N_2) / f_i(N_1) f_j(N_2)$$

$$f_i(A) = 0.5$$

$$f_j(A) = 0.5$$

$$f_i(U) = 0.5$$

$$f_j(U) = 0.5$$

$$f_{ij}(A:U) = 0.5$$

$$f_{ij}(U:A) = 0.5$$

$$f_{ij}(X:Y) = 0.0$$

$$H_{ij} = 0.5 \log_2 0.5 / 0.5^2 + 0.5 \log_2 0.5 / 0.5^2 = 0.5 + 0.5 = 1$$

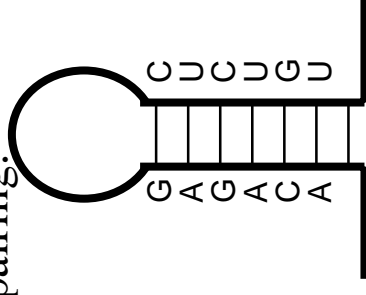
(Note: Max is 1 due to considering only A and U)

RNA Structure

Phylogeny based prediction

For a random sequence, i.e. one with no indication of base-pairing:

...	A	...	U	...
...	A	...	A	...
...	U	...	U	...
...	U	...	A	...
	<i>i</i>			<i>j</i>



$$H(i,j) = \sum f_{ij}(N_1, N_2) \log_2 f_{ij}(N_1, N_2) / f_i(N_1) f_j(N_2)$$

$$f_i(A) = 0.5 \quad f_j(A) = 0.5$$

$$f_i(U) = 0.5 \quad f_j(U) = 0.5$$

$$f_{ij}(A:U) = 0.25 \quad f_{ij}(U:A) = 0.25 \quad f_{ij}(X:Y) = 0.0$$

$$\begin{aligned} H_{ij} &= 0.5 \log_2 0.25 / 0.5^2 + 0.5 \log_2 0.25 / 0.5^2 \\ &= 0.0 + 0.0 = 0 \end{aligned}$$

RNA Structure

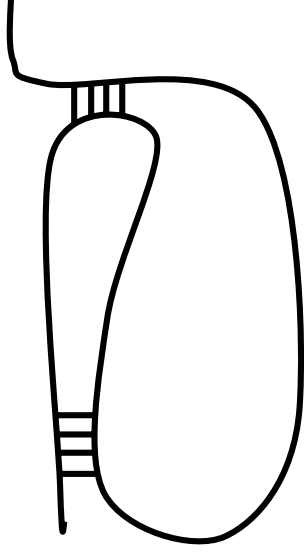
Phylogeny based prediction

- Inference of structure from covariance or mutual information depends on having the correct alignment
- Correct alignment depends on knowing the correct structures
- Can only find common structures, not structures unique to a molecule
- Can, in principle, detect pseudoknots

RNA Structure

Pseudonots

- Pseudoknots are very important for biological function



- Predicting pseudoknots
 - Requires additional N^2 time, but theoretically possible
 - Most approaches use some heuristic search

RNA Structure

Pseudoknots

