RNA Secondary Structure Prediction

Introduction to RNA Sequence/Structure Analysis

- RNAs have many structural and functional uses
  - Translation
  - Transcription
  - RNA splicing
  - RNA processing and editing
  - cellular localization
  - catalysis

RNA functions

- RNA functions as
  - mRNA
  - rRNA
  - tRNA
  - In nuclear export
  - Part of spliceosome: (snRNA)
  - Regulatory molecules (RNAi)
  - Enzymes
  - Viral genomes
  - Retrotransposons
  - Medicine

Biological Functions of Nucleic Acids

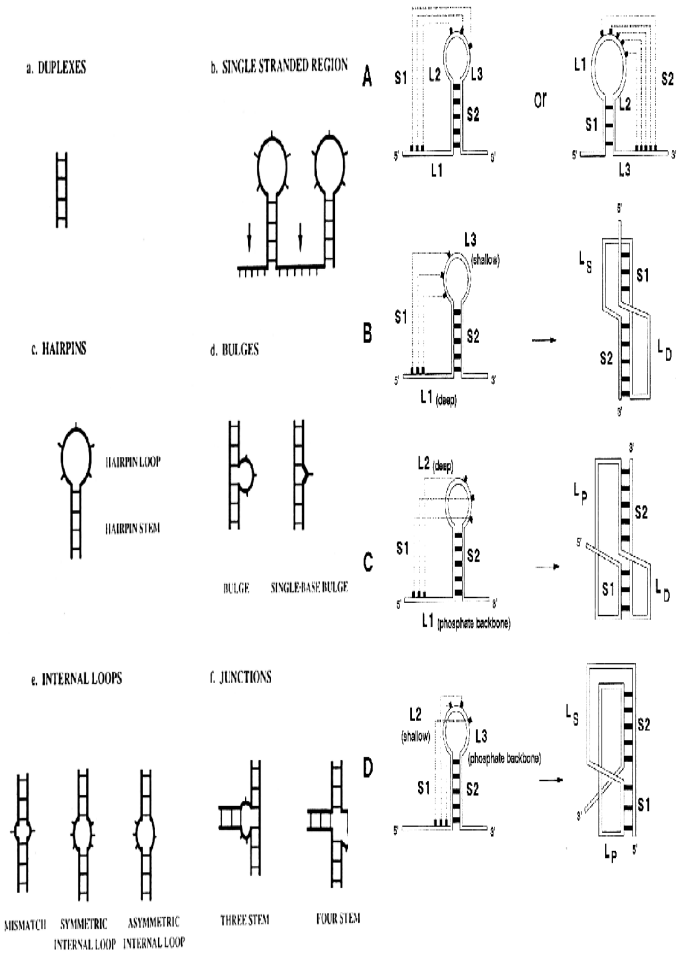
- tRNA (transfer RNA, adaptor in translation)
- rRNA (ribosomal RNA, component of ribosome)
- snRNA (small nuclear RNA, component of splicesome)
- snoRNA (small nucleolar RNA, takes part in processing of rRNA)
- RNase P (ribozyme, processes tRNA)
- SRP RNA (RNA component of signal recognition particle)
- .....

RNA Sequence Analysis

- RNA sequence analysis different from DNA sequence analysis
- RNA structures fold and base pair to form secondary structures
- not necessarily the sequence but structure conservation is most important with RNA

Secondary Structures of Nucleic Acids

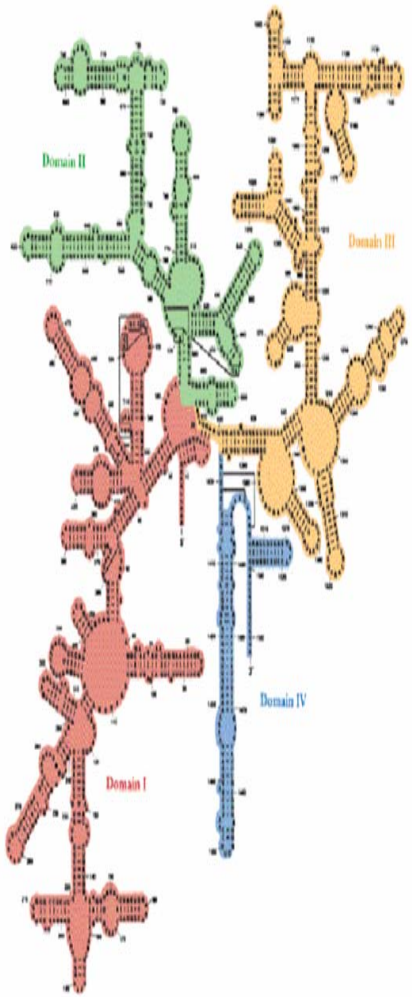
- DNA is primarily in duplex form.
- RNA is normally single stranded which can have a diverse form of secondary structures other than duplex.



Source: Cornelis W. A. Pleij in Gesteland, R. F. and Atkins, J. F. (1993) THE RNA WORLD. Cold Spring Harbor Laboratory Press.

More Secondary Structures

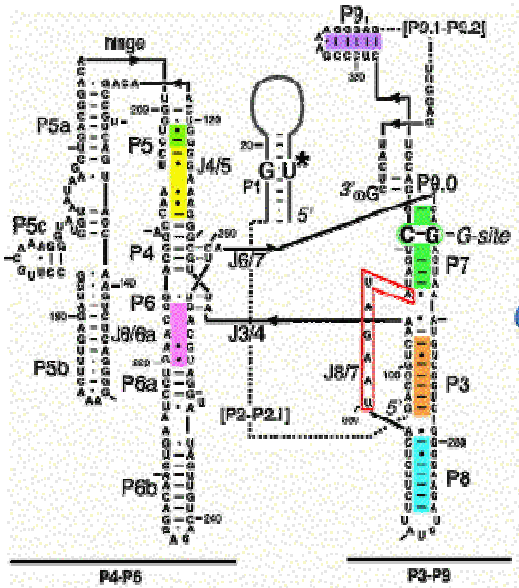
Pseudoknots



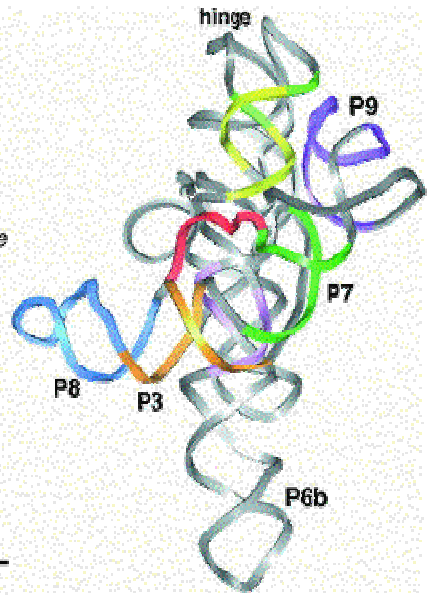
rRNA Secondary Structure Based on **Phylogenetic** Data

3D Structures of RNA: Catalytic RNA

Secondary Structure
Of Self-splicing RNA



Tertiary Structure
Of Self-splicing RNA



Some structural rules:

- Base pairing is stabilizing
- Unpaired sections (loops) destabilize
- 3D conformation with interactions makes up for this

RNA secondary structure

- E. coli Rnase P RNA secondary structure

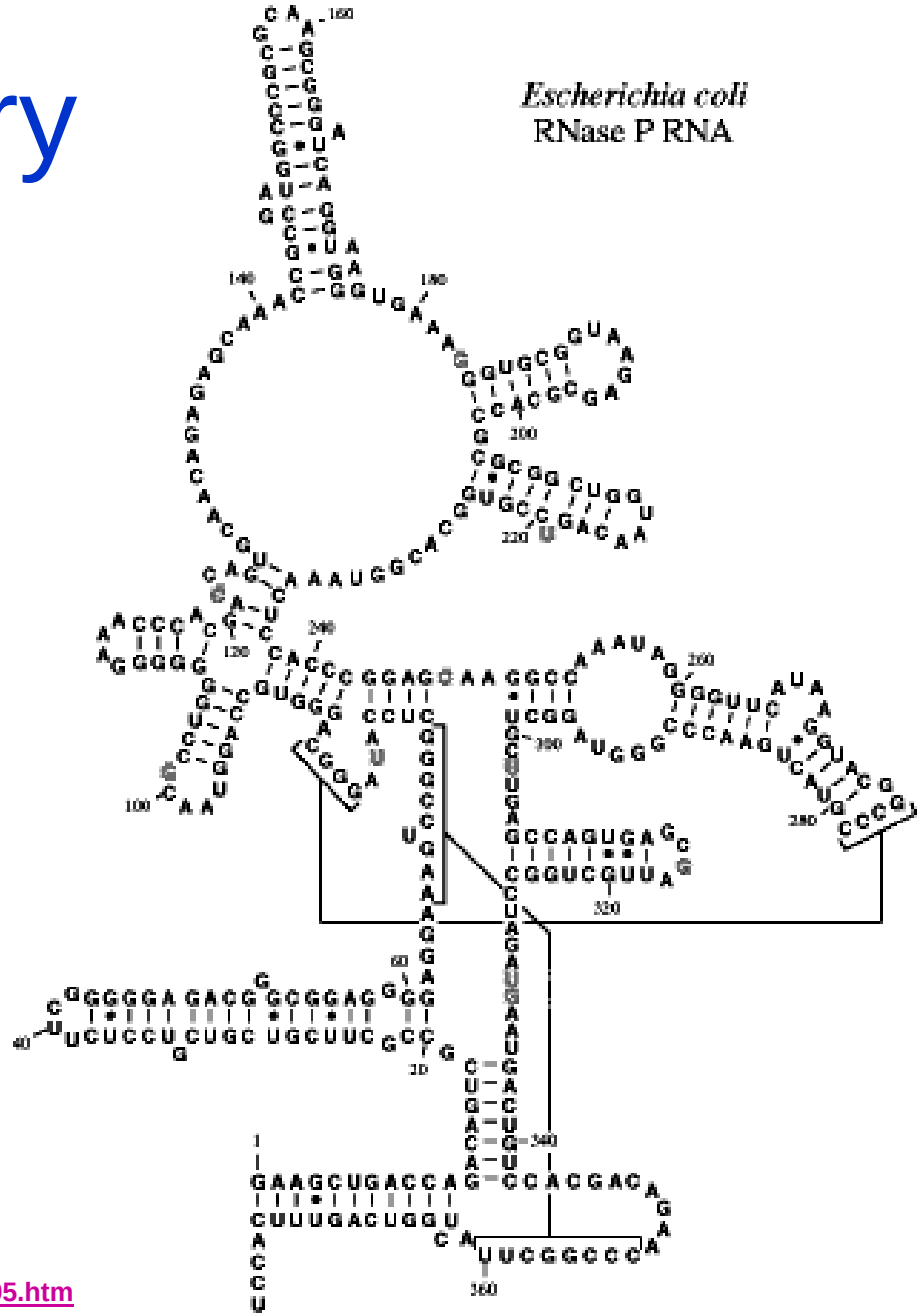
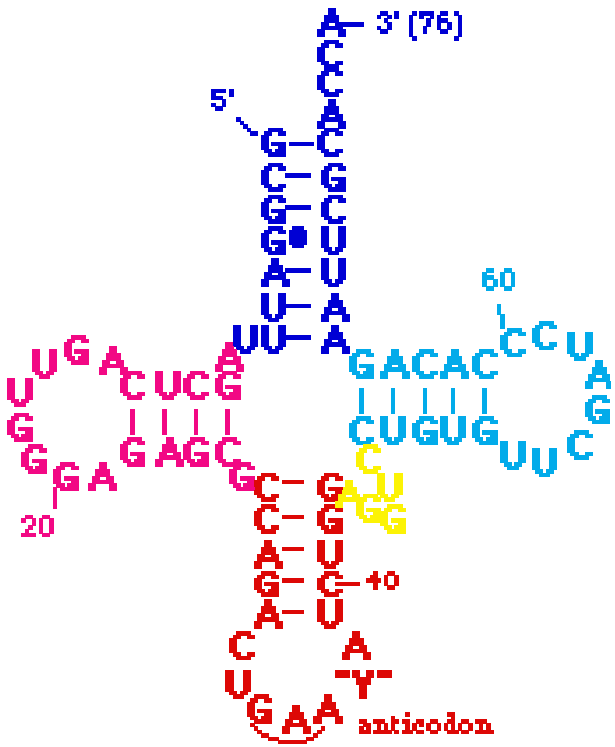
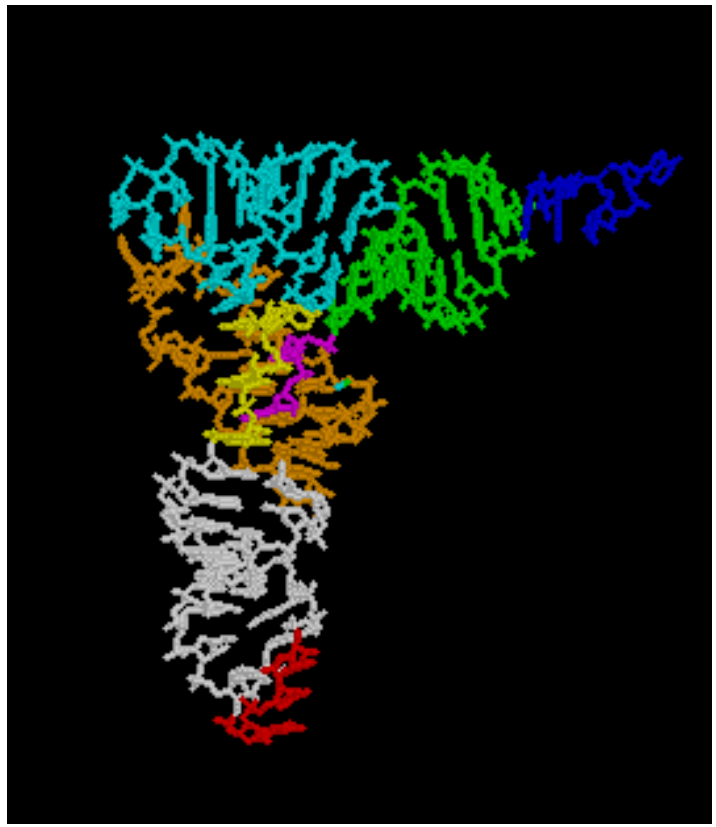


Image source: www.mbio.ncsu.edu/JWB/MB409/lecture/ lecture05/lecture05.htm

tRNA structure





Features of RNA

- RNA: polymer composed of a combination of four nucleotides
  - adenine (A)
  - cytosine (C)
  - guanine (G)
  - uracil (U)

Features of RNA

- G-C and A-U form complementary hydrogen bonded base pairs (canonical Watson-Crick)
- G-C base pairs being more stable (3 hydrogen bonds) A-U base pairs less stable (2 bonds)
- non-canonical pairs can occur in RNA -- most common is G-U

Features of RNA

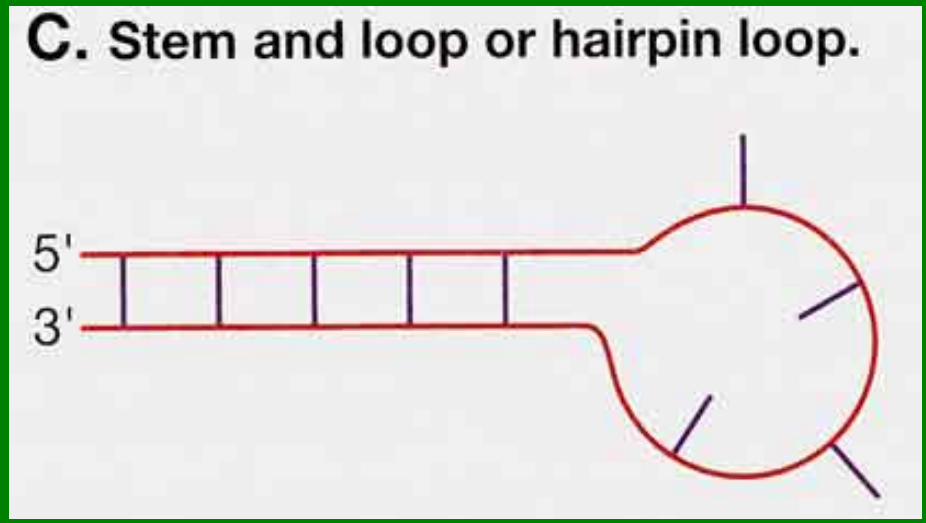
- RNA typically produced as a single stranded molecule (unlike DNA)
- Strand folds upon itself to form base pairs
- secondary structure of the RNA

Features of RNA

- intermediary between a linear molecule and a three-dimensional structure
- Secondary structure mainly composed of double-stranded RNA regions formed by folding the single-stranded RNA molecule back on itself

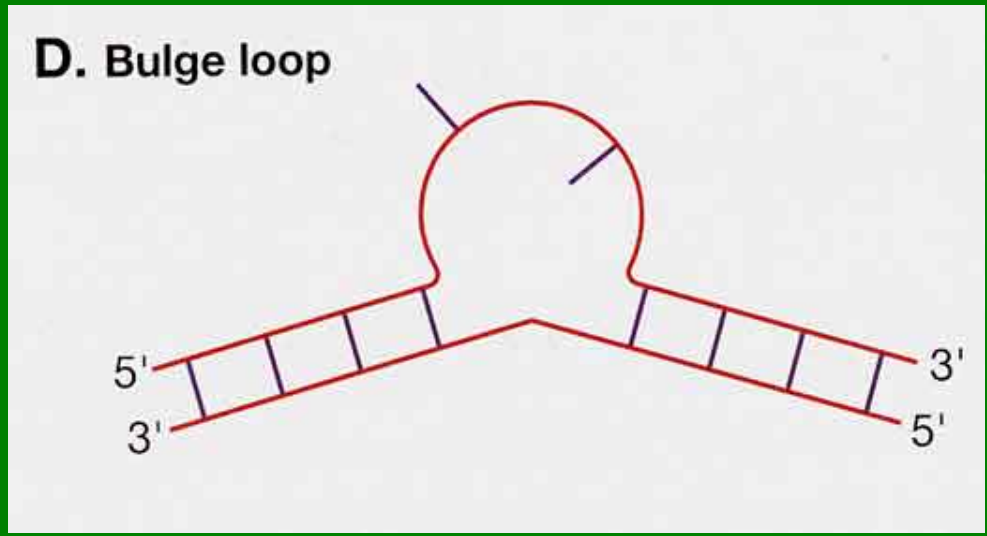
Stem Loops (Hairpins)

- Loops generally at least 4 bases long



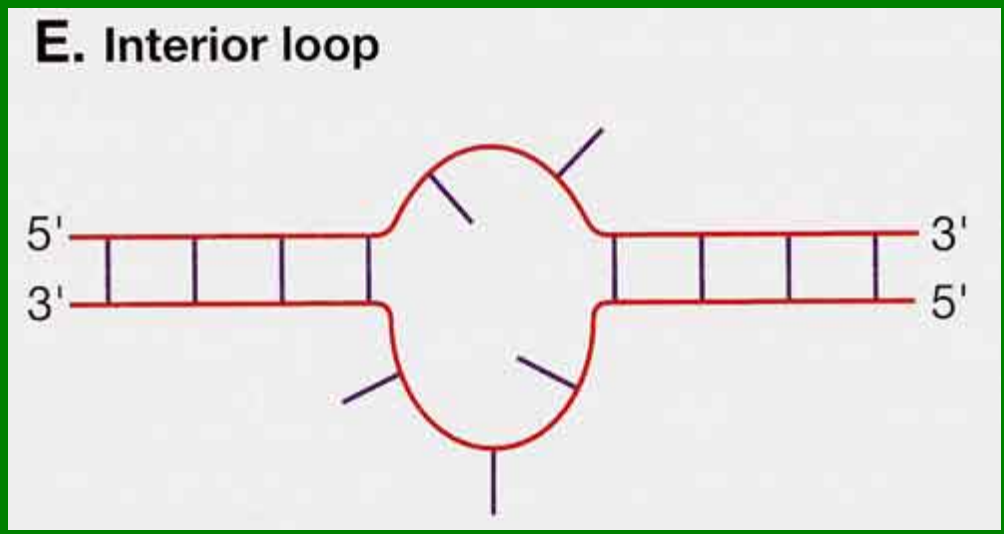
Bulge Loops

- occur when bases on one side of the structure cannot form base pairs



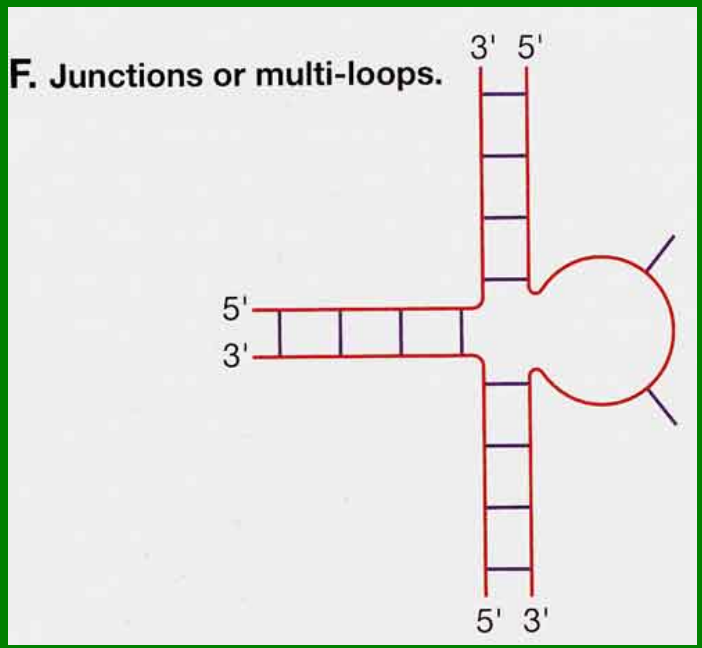
Interior Loops

- occur when bases on both sides of the structure cannot form base pairs



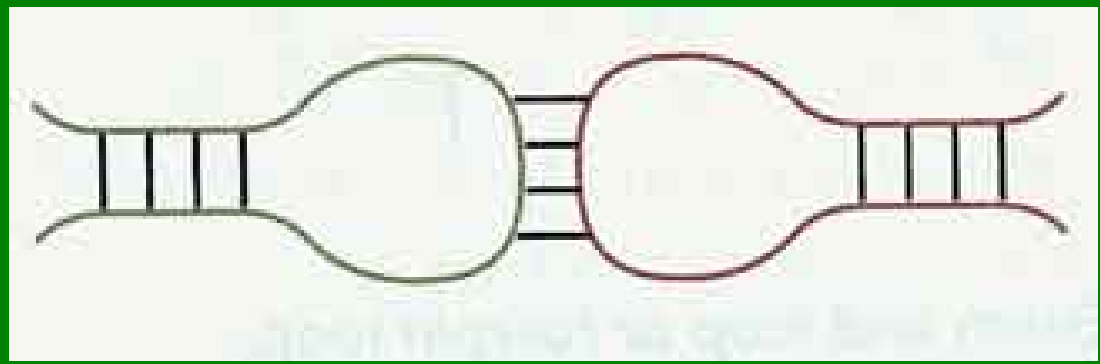
Junctions (Multiloops)

- two or more double-stranded regions converge to form a closed structure

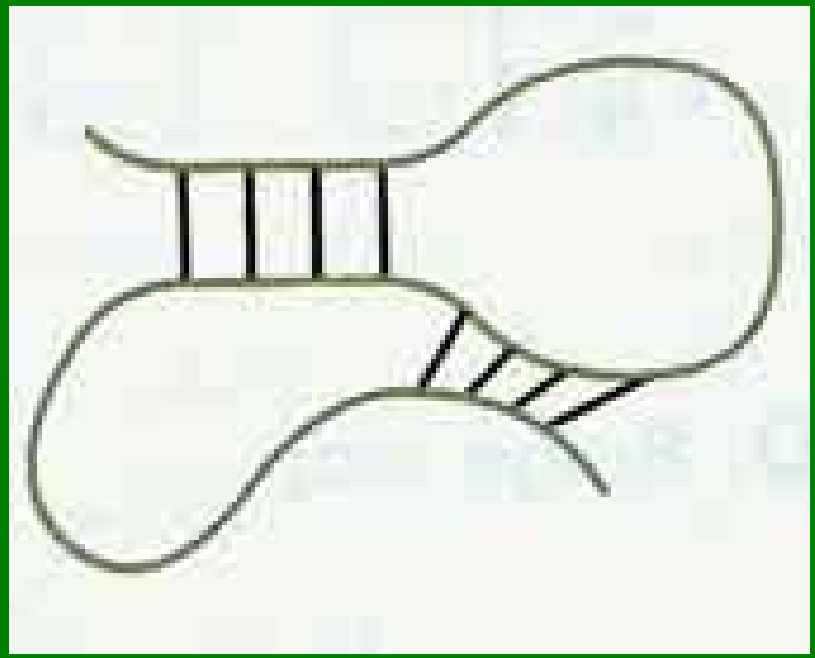


Kissing Hairpins

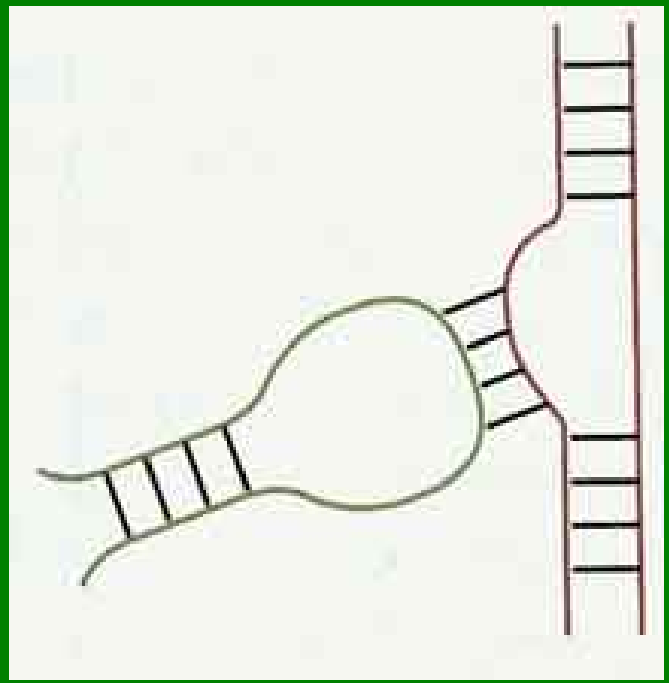
- unpaired bases of two separate hairpin loops base pair with one another



Pseudoknots



Hairpin-Bulge Interactions



RNA structure prediction methods

- Dot Plot Analysis
- Base-Pair Maximization
- Free Energy Methods
- Covariance Models

How RNA Prediction Methods Were Developed

- Nussinov and Jacobson (1980), Zuker and Stiegler (1981), Trifonov and Bolshoi (1983) ....

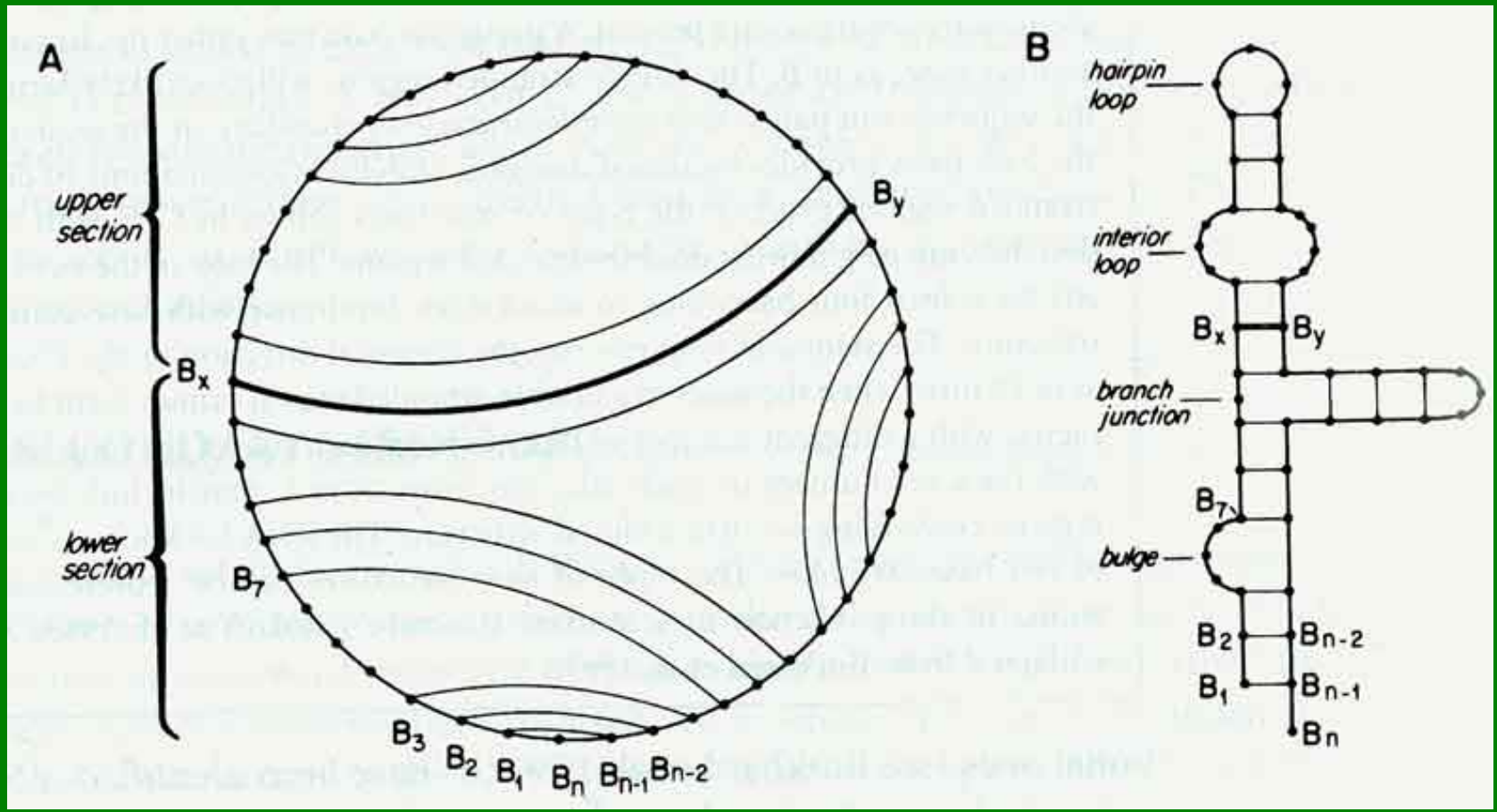
Main approaches to RNA secondary structure prediction

- Energy minimization
  - dynamic programming approach
  - does not require prior sequence alignment
  - require estimation of energy terms contributing to secondary structure
- Comparative sequence analysis
  - Using sequence alignment to find conserved residues and covariant base pairs.
  - most trusted

Circular Representation

- base pairs of a secondary structure represented by a circle
- arc drawn for each base pairing in the structure
- If any arcs cross, a pseudoknot is present

Circular Representation



Base-Pair Maximization

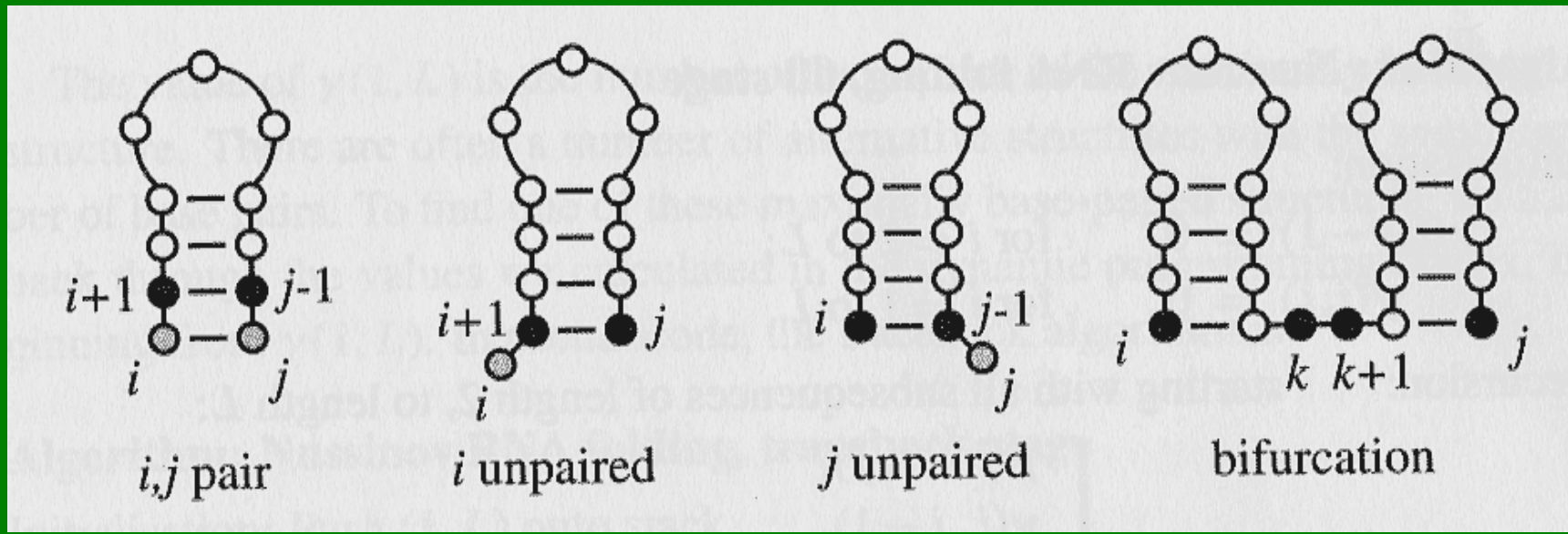
- Find structure with the most base pairs
- Efficient dynamic programming approach to this problem introduced by Ruth Nussinov (Tel-Aviv, 1970s).

Nussinov Algorithm

- Four ways to get the best structure between position i and j from the best structures of the smaller subsequences

1) Add i,j pair onto best structure found for subsequence i+1, j-1
2) add unpaired position i onto best structure for subsequence i+1, j
3) add unpaired position j onto best structure for subsequence i, j-1
4) combine two optimal structures i,k and k+1, j

Nussinov Algorithm



Nussinov Algorithm

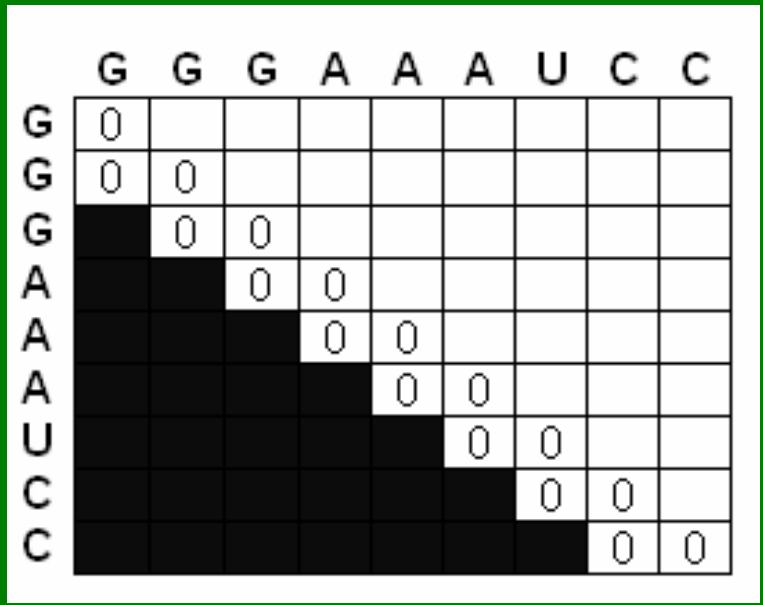
- compares a sequence against itself in a dynamic programming matrix
-
- Four rules for scoring the structure at a particular point
- Since structure folds upon itself, only necessary to calculate half the matrix

Nussinov Algorithm

- Initialization: score for matches along main diagonal and diagonal just below it are set to zero

- Formally, the scoring matrix, M, is initialized:
  - M[i][i] = 0 for i = 1 to L (L is sequence length)
  - M[i][i-1] = 0 for i = 2 to L

Nussinov Algorithm

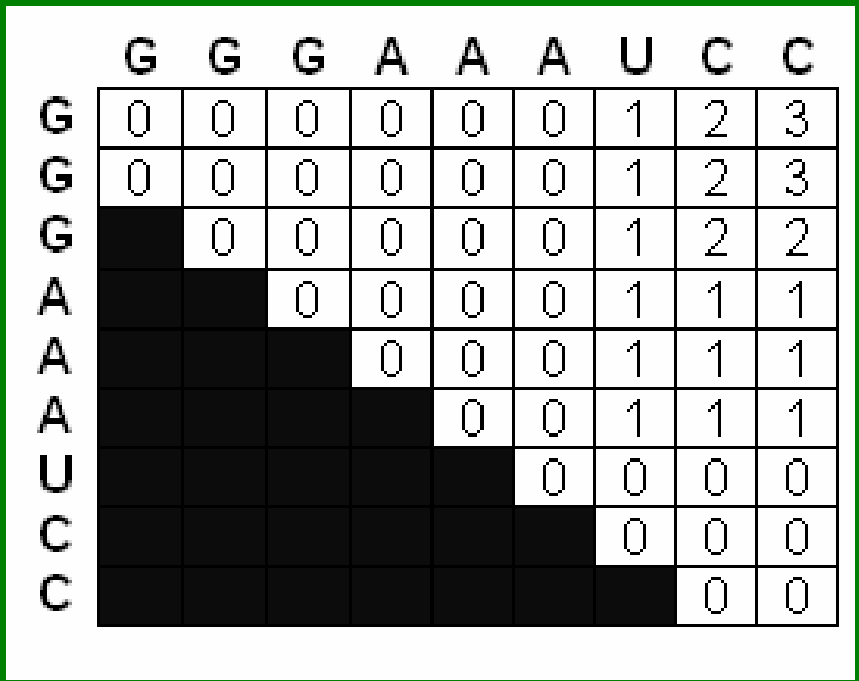


Nussinov Algorithm

- Matrix Fill:

- M[i][j] = max of the following :
  - M[i+1][j] (*ith residue is hanging off by itself*)
  - M[i][j-1] (*jth residue is hanging off by itself*)
  - M[i+1][j-1] + $S(x_i, x_j)$ (*ith and jth residue are paired; if x_i = complement of x_j, then $S(x_i, x_j) = 1$; otherwise it is 0.*)
  - M[i][j] = $MAX_{i<k<j}$ (M[i][k] + M[k+1][j]) (*merging two substructures*)

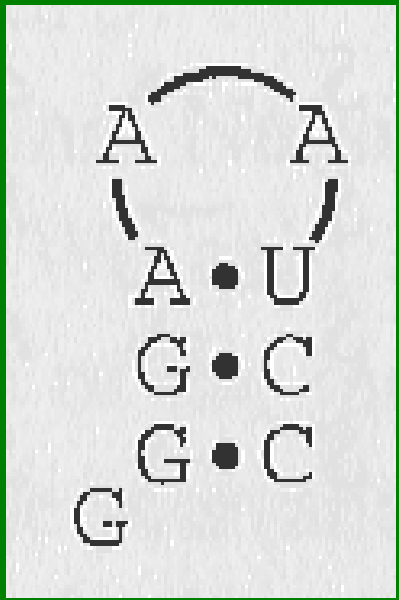
Nussinov Algorithm



	G	G	G	A	A	A	U	C	C
G	0	0	0	0	0	0	1	2	3
G	0	0	0	0	0	0	1	2	3
G		0	0	0	0	0	1	2	2
A			0	0	0	0	1	1	1
A				0	0	0	1	1	1
A					0	0	1	1	1
U						0	0	0	0
C							0	0	0
C								0	0

Nussinov Algorithm

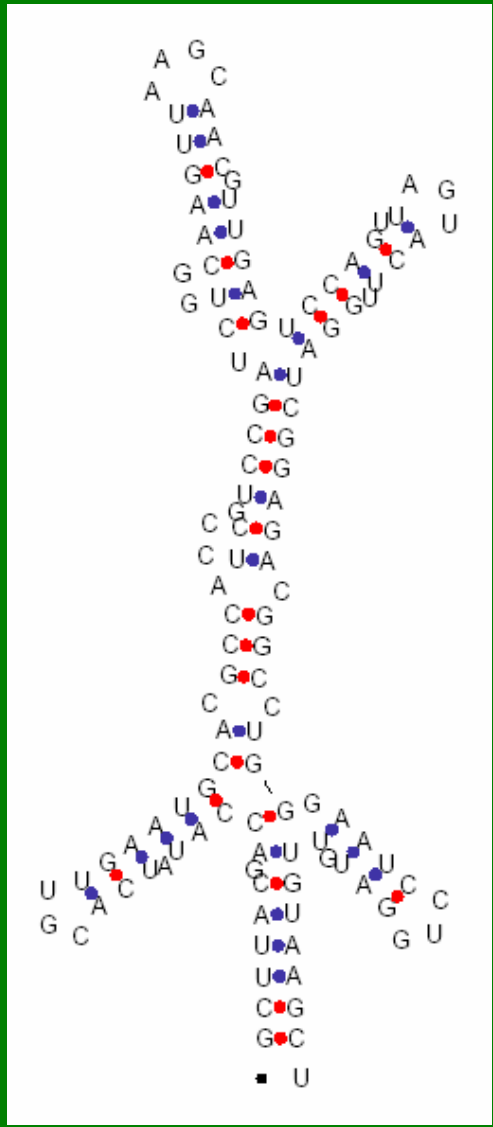
- Traceback (P 271, Durbin et al) leads to the following structure:



Nussinov Algorithm

http://ludwig-sun2.unil.ch/~bsondere/nussinov/

Nussinov Results



Evaluation of Maximizing Basepairs

- Simplistic approach
- Does not give accurate structure predictions.
- Misses:
  - nearest neighbor interactions
  - stacking interactions
  - loop length preferences

Free Energy Minimization RNA Structure Prediction

- All possible choices of complementary sequences are considered
- Set(s) providing the most energetically stable molecules are chosen
- When RNA is folded, some bases are paired with other while others remain free, forming “loops” in the molecule.
- Speaking qualitatively, bases that are bonded tend to stabilize the RNA (i.e., have negative free energy), whereas unpaired bases form destabilizing loops (positive free energy).
- Through thermodynamics experiments, it has been possible to estimate the free energy of some of the common types of loops that arise.
- Because the secondary structure is related to the function of the RNA, we would like to be able to predict the secondary structure.
- Given an RNA sequence, the *RNA Folding Problem* is to predict the secondary structure that minimizes the total free energy of the folded RNA molecule.

Prediction of Minimum-Energy RNA Structure is Limited

- In predicting minimum energy RNA secondary structure, several simplifying assumptions are made.
  - The most likely structure is identical to the energetically preferable structure
  - Nearest-neighbor energy calculations give reliable estimates of an experimentally achievable energy measurements
  - Usually we can neglect pseudoknots

Assumptions in secondary Structure Prediction

- most likely structure similar to energetically most stable structure
- Energy associated with any position is only influenced by local sequence and structure

  Structure formed does not produce pseudoknots

Inferring Structure By Comparative Sequence Analysis

- most reliable computational method for determining RNA secondary structure

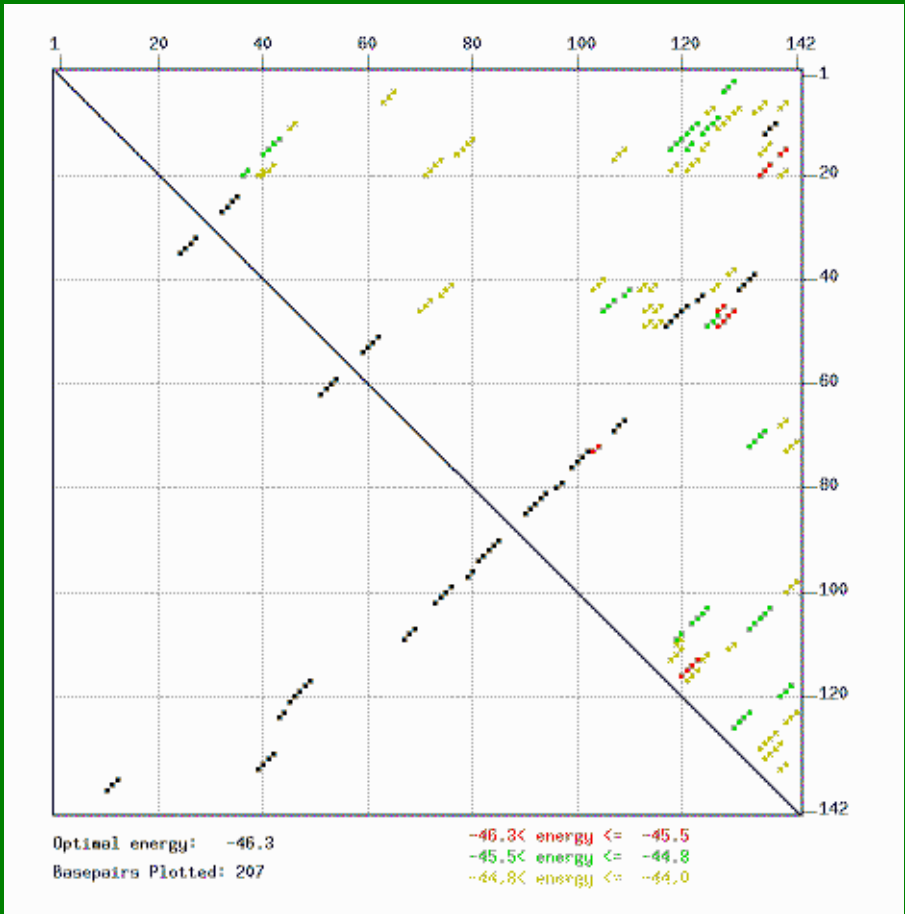
- consider the example from Durbin, et al., p 266
- See an additional lecture of David Mathews

Predicting Structure From a Single Sequence

- RNA molecule only 200 bases long has 10^{50} possible secondary structures

- Find self-complementary regions in an RNA sequence using a dot-plot of the sequence against its complement
  - repeat regions can potentially base pair to form secondary structures
  - advanced dot-plot techniques incorporate free energy measures

Dot Plot



- Image Source: http://www.finchcms.edu/cms/biochem/Walters/rna_folding.html

Energy Minimization Methods

- RNA folding is determined by biophysical properties
- Energy minimization algorithm predicts the correct secondary structure by minimizing the free energy (ΔG)
- ΔG calculated as sum of individual contributions of:
  - loops
  - base pairs
  - secondary structure elements
- Energies of stems calculated as stacking contributions between neighboring base pairs

Energy Minimization Methods

- Free-energy values (kcal/mole at 37^{o}C) are as follows:

	Stacking Energies for base pairs					
	A/U	C/G	G/C	U/A	G/U	U/G
A/U	-0.9	-1.8	-2.3	-1.1	-1.1	-0.8
C/G	-1.7	-2.9	-3.4	-2.3	-2.1	-1.4
G/C	-2.1	-2.0	-2.9	-1.8	-1.9	-1.2
U/A	-0.9	-1.7	-2.1	-0.9	-1.0	-0.5
G/U	-0.5	-1.2	-1.4	-0.8	-0.4	-0.2
U/G	-1.0	-1.9	-2.1	-1.1	-1.5	-0.4

Energy Minimization Methods

- Free-energy values (kcal/mole at 37^{o}C) are as follows:

	Destabilizing Energies for Loops				
Number of Bases	1	5	10	20	30
Internal	--	5.3	6.6	7.0	7.4
Bulge	3.9	4.8	5.5	6.3	6.7
Hairpin	--	4.4	5.3	6.1	6.5

Energy Minimization Methods

- Given the energy tables, and a folding, the free energy can be calculated for a structure

Calculating Best Structure

- sequence is compared against itself using a dynamic programming approach
  - similar to the maximum base-paired structure

- instead of using a scoring scheme, the score is based upon the free energy values
- Gaps represent some form of a loop
- The most widely used software that incorporates this minimum free energy algorithm is MFOLD.

Free Energy Minimization RNA Structure Prediction

- http://www.bioinfo.rpi.edu/~zukerm/Bio-5495/RNAfold-html/

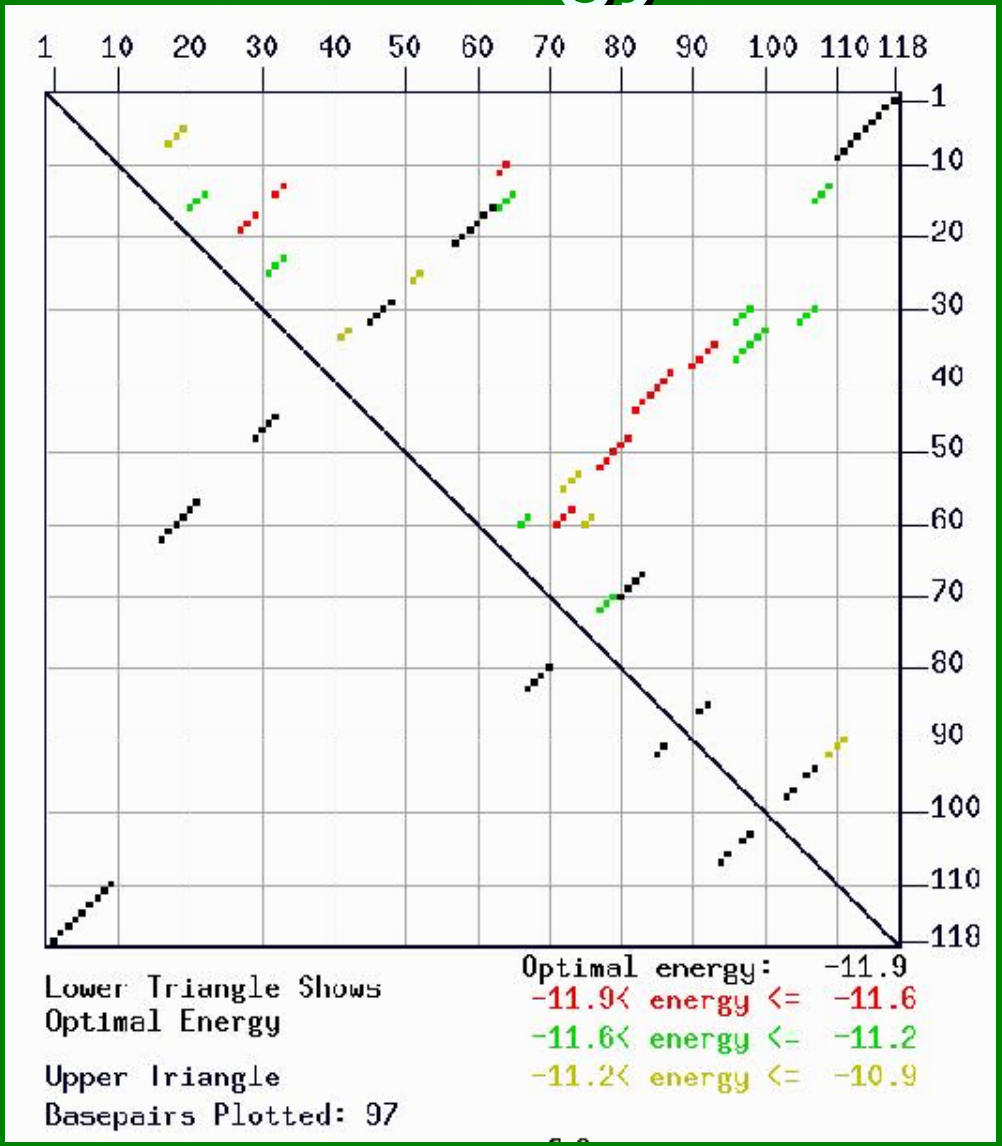
Calculating Best Structure

- most widely used software incorporating minimum free energy algorithm is MFOLD
- http://www.bioinfo.rpi.edu/applications/mfold/
- http://www.bioinfo.rpi.edu/applications/mfold/old/rna/

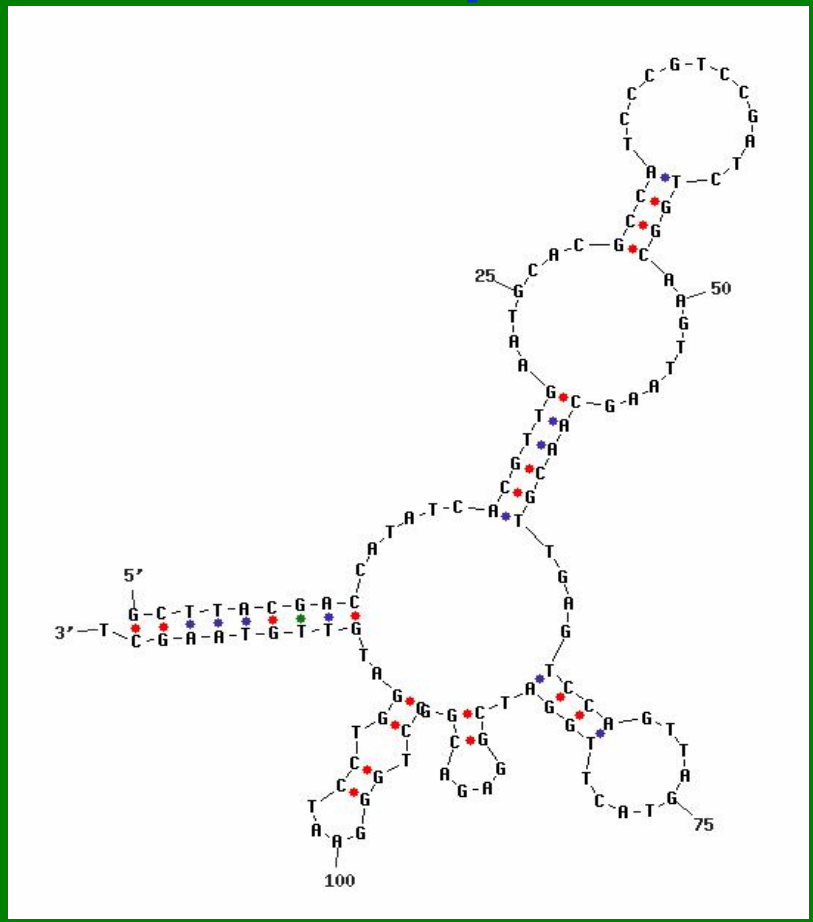
Example Sequence

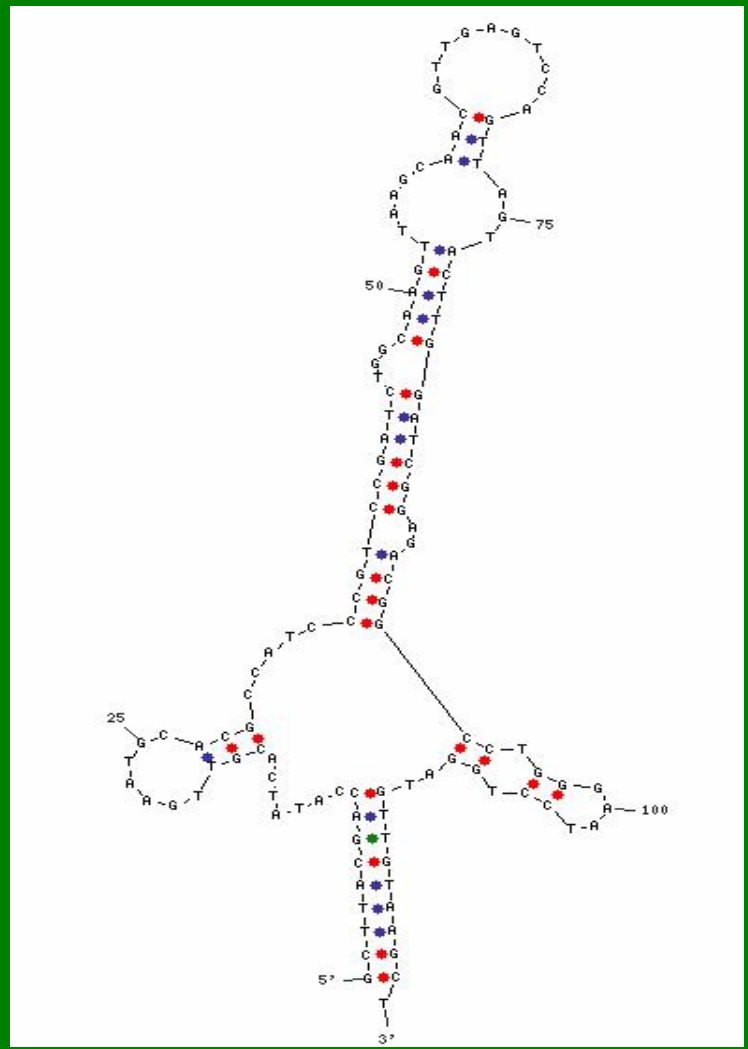
GCTTACGACCATATCACGTTGAATGCACGC
CATCCCGTCCGATCTGGCAAGTTAAGCAAC
GTTGAGTCCAGTTAGTACTTGGATCGGAGA
CGGCCTGGGAATCCTGGATGTTGTAAGCT

MFOLD Energy Dot Plot



Optimal Structure

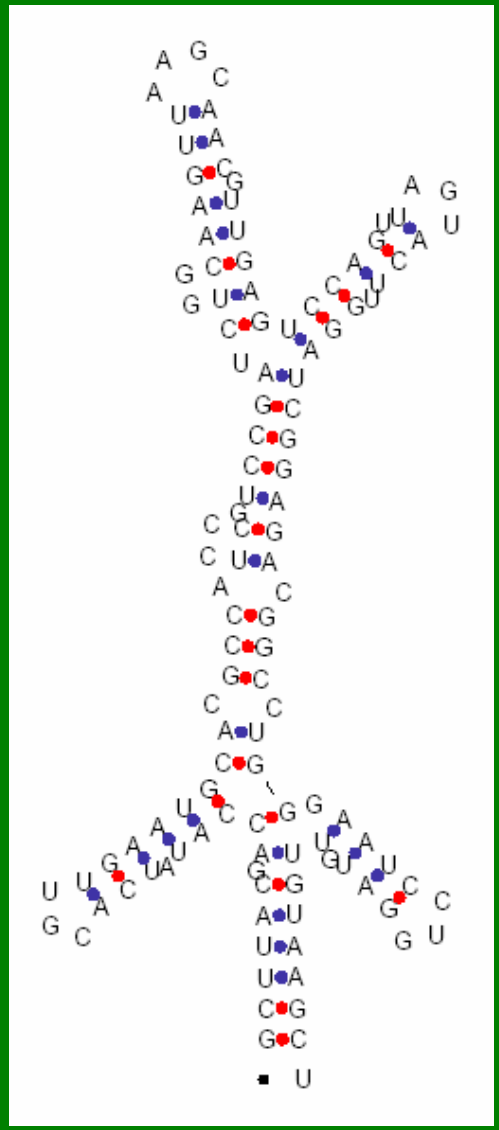


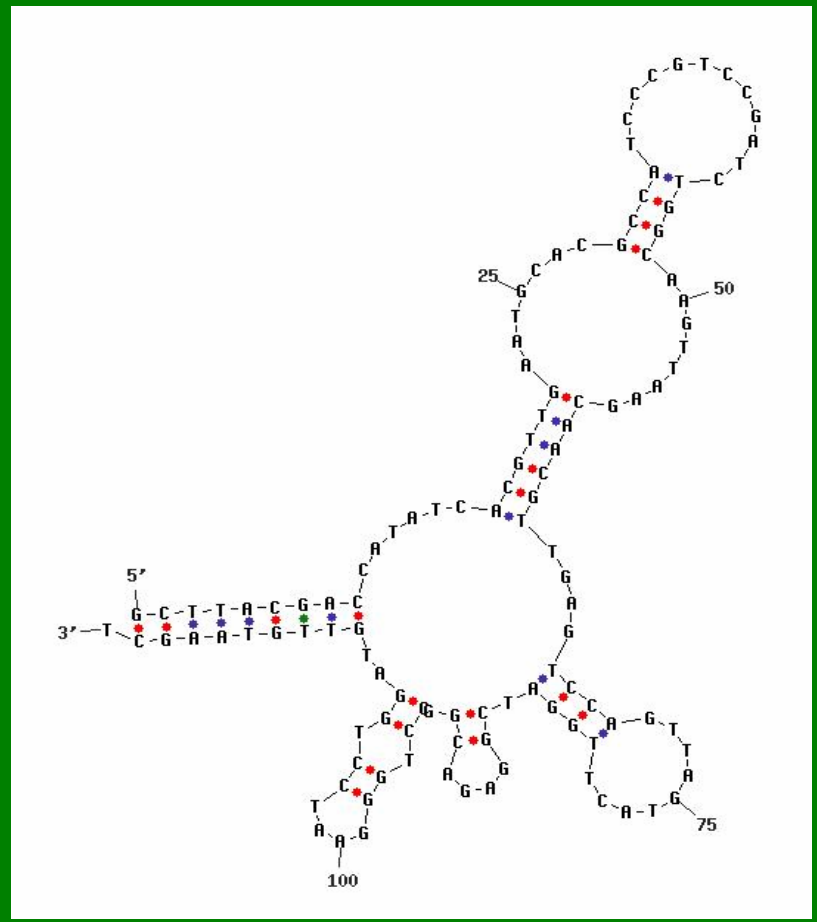


Suboptimal Folds

- The correct structure is not necessarily structure with optimal free energy
- within a certain threshold of the calculated minimum energy
- MFOLD updated to report suboptimal folds

Comparison of Methods





Open Problem: Pseudoknots.

Example of a partial solution:
Rivas and Eddy algorithm

- running time is $O(n^6)$
- "*we lack a systematic a priori characterization of the class of configurations that this algorithm can solve*" (Rivas and Eddy, 1999)