RNA folding

RNA secondary structure prediction by dynamic programming algorithms

Yuzhen Ye

School of Informatics

Indiana University

RNAs have diverse functions

- Protein synthesis (rRNA and tRNA)
- RNA processing (snoRNA)
- Gene regulation
  - RNA interference (RNAi)
  - Andrew Fire and Craig Mello (2006 Nobel prize)
- DNA-like function
  - Virus
- RNA world

Riboswitch

- What’s riboswitch
- Riboswitch mechanism

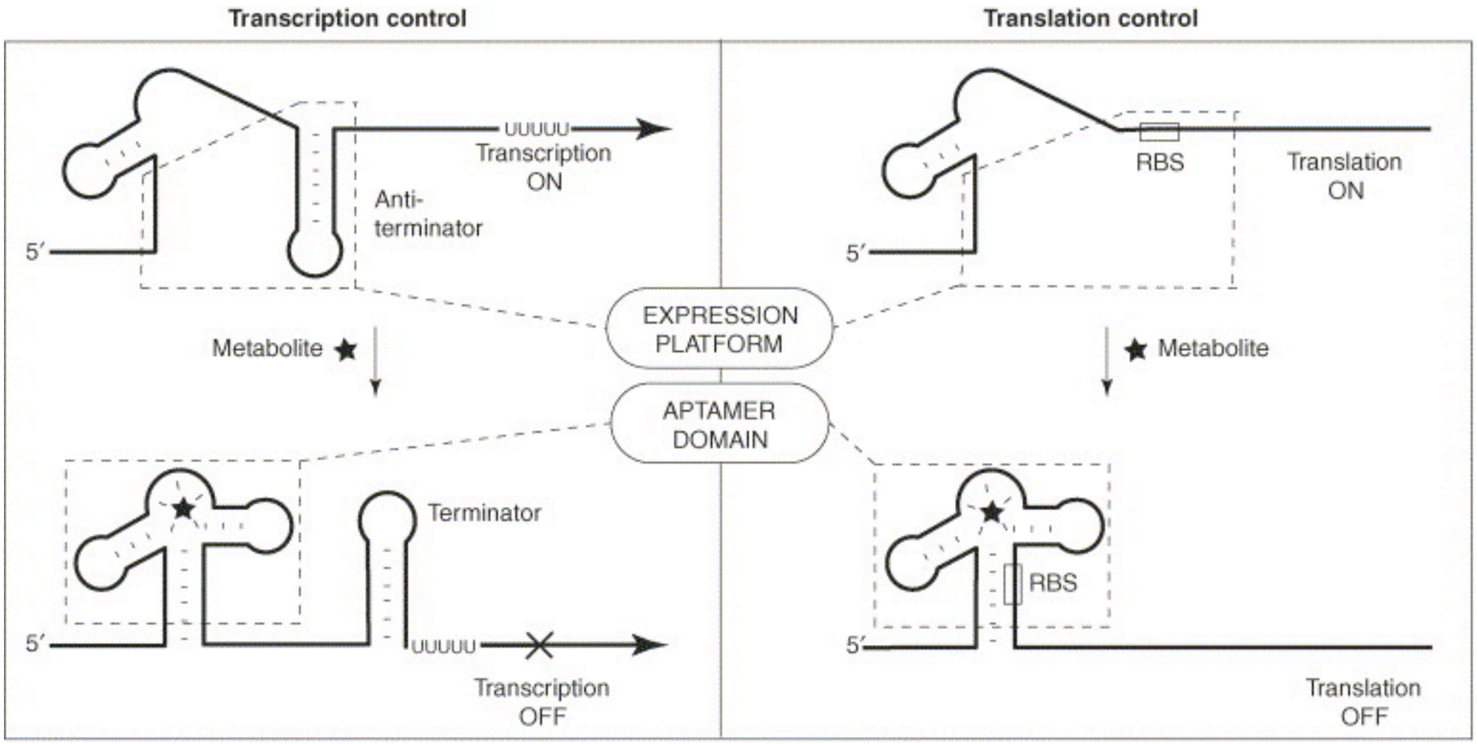
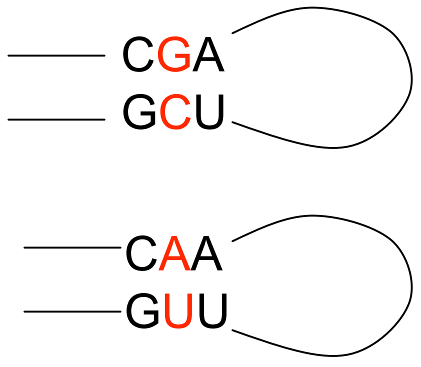


Image source: Curr Opin Struct Biol. 2005, 15(3):342-348

Structures are more conserved

- Structure information is important for alignment (and therefore gene finding)

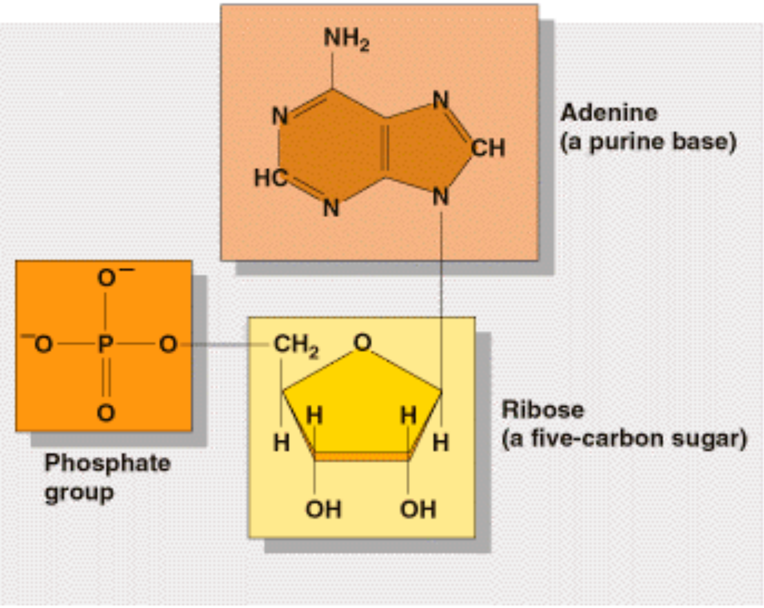


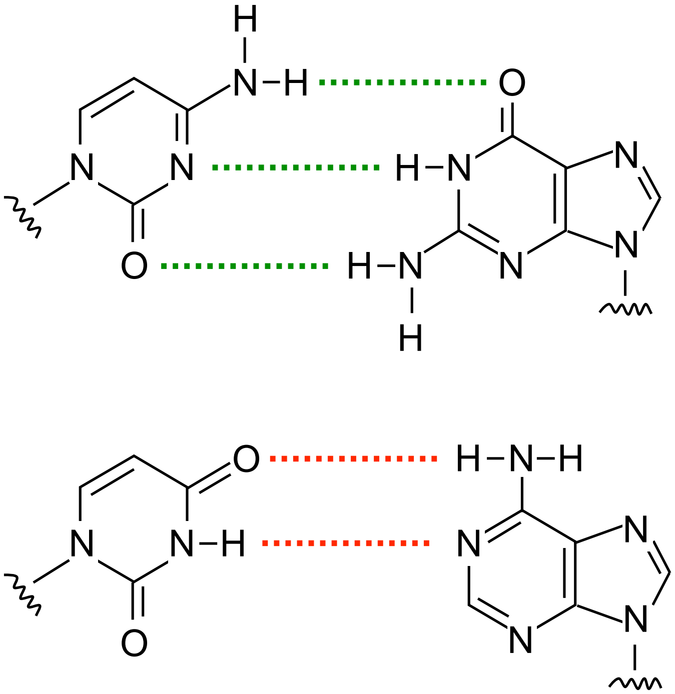
Features of RNA

- RNA typically produced as a single stranded molecule (unlike DNA)
- Strand folds upon itself to form base pairs & secondary structures
- Structure conservation is important

- RNA sequence analysis is different from DNA sequence

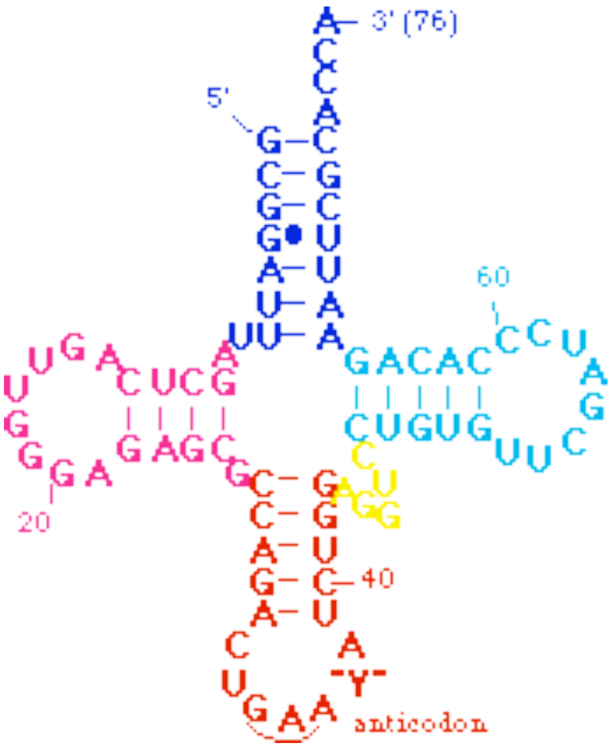
Canonical base pairing

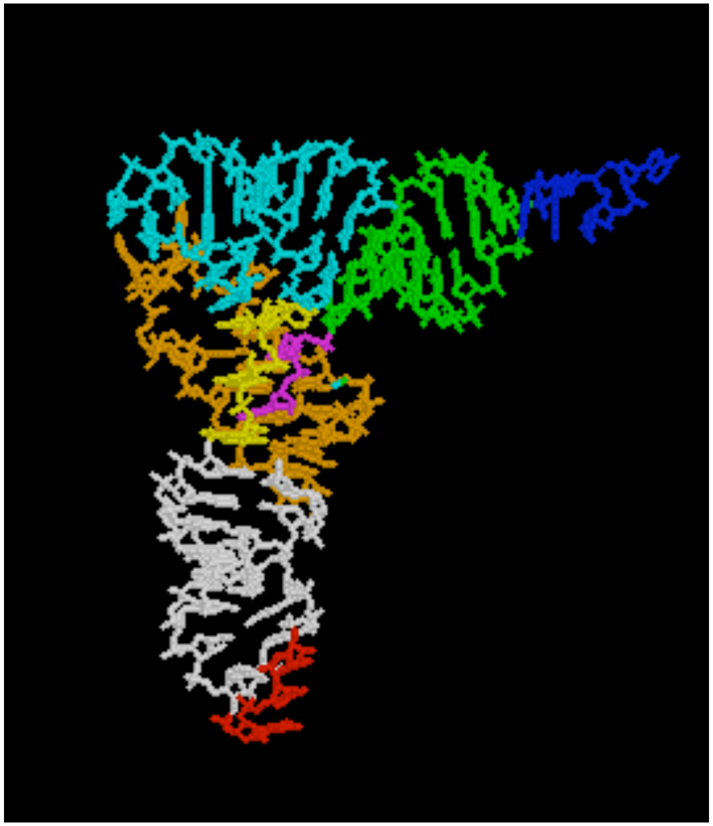




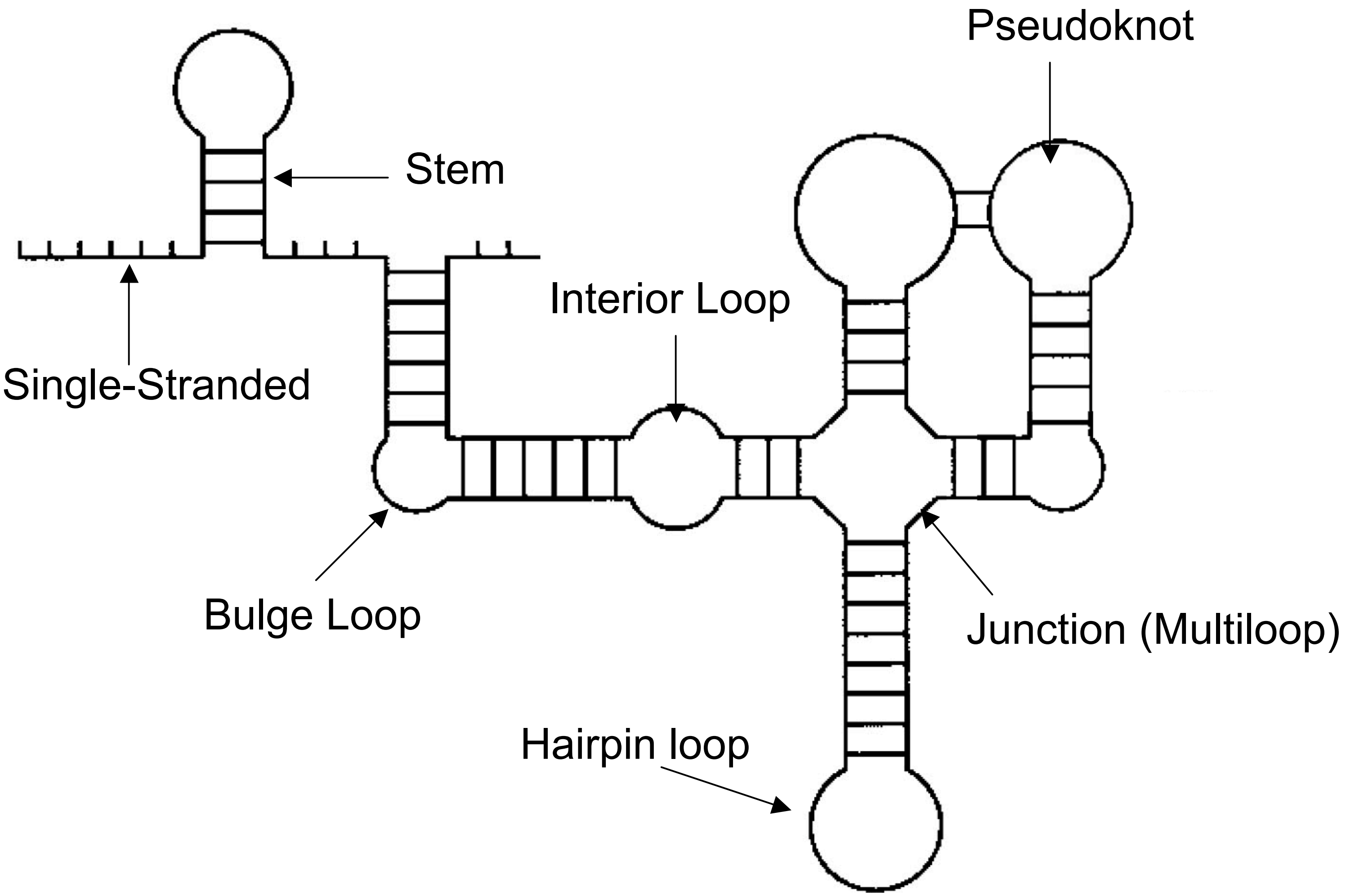
Watson-Crick base pairing
Non-Watson-Crick base pairing G/U (Wobble)

tRNA structure

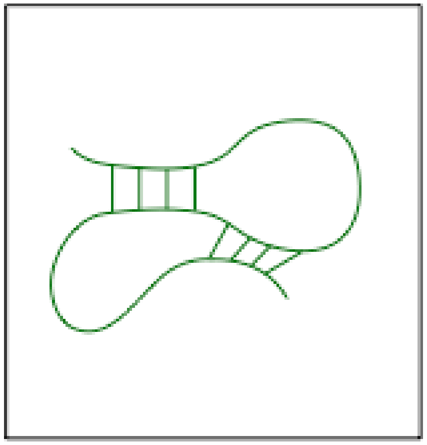




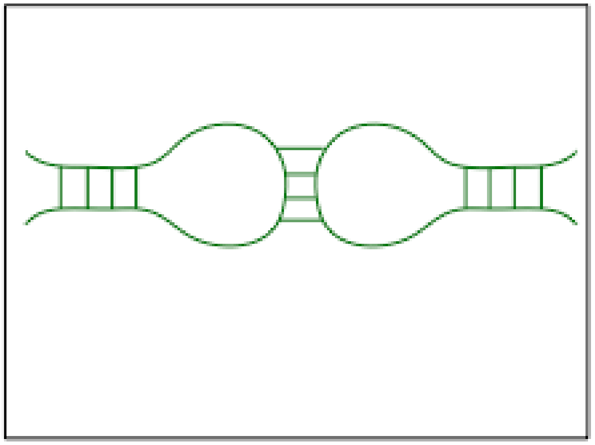
RNA secondary structure



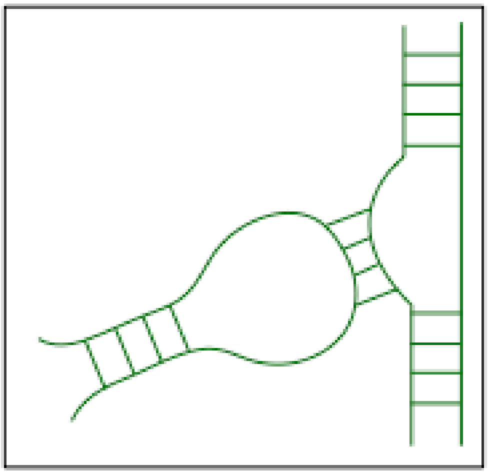
Complex folds



Pseudoknot

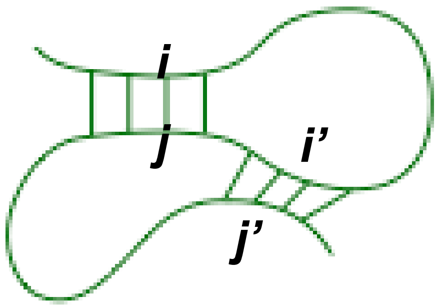


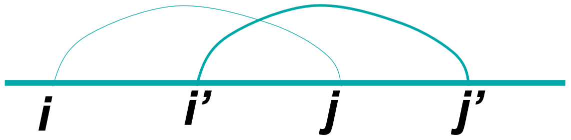
Kissing Hairpins



Hair-bulge interaction

Pseudoknots





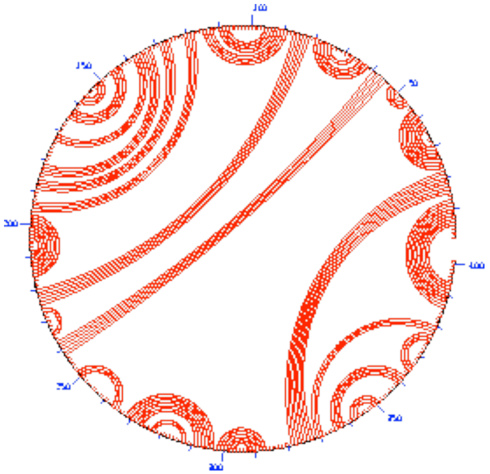
$$i < i' < j < j'$$

? $i < i' < j' < j$

$i < j < i' < j'$

RNA secondary structure representation

- 2D
- Circle plot
- Dot plot
- Mountain
- Parentheses
- Tree model



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

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
- Simultaneous folding and alignment (structural alignment)

Assumptions in energy minimization approaches

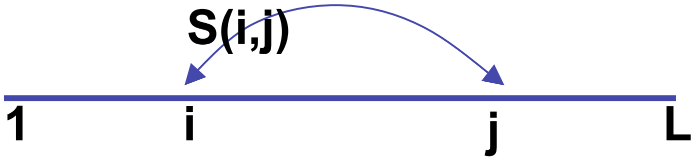
- Most likely structure similar to energetically most stable structure
- Energy associated with any position is only influenced by local sequence and structure
- Neglect pseudoknots

Base-pair maximization

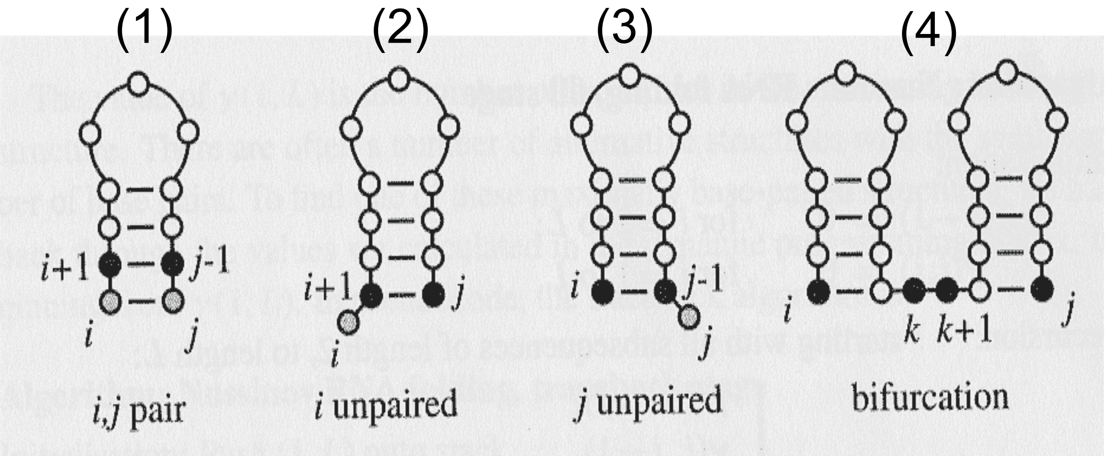
- Find structure with the most base pairs
  - Only consider A-U and G-C and do not distinguish them
- Nussinov algorithm (1970s)
  - Too simple to be accurate, but stepping-stone for later algorithms

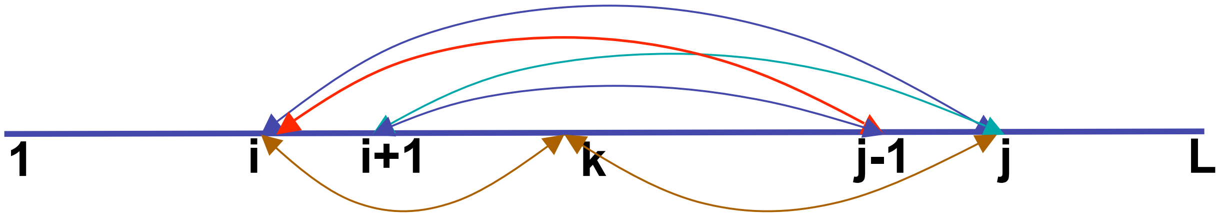
Nussinov algorithm

- Problem definition
  - Given sequence $X=x_1x_2\ldots x_{L,}$ compute a structure that has maximum (weighted) number of base pairings
- How can we solve this problem?
  - Remember: RNA folds back to itself!
  - S(i,j) is the maximum score when $x_i..x_j$ folds optimally
  - S(1,L)?
  - S(i,i)?



“Grow” from substructures





$$S(i,j) = max \begin{cases} S(i+1, j-1) + w(i,j) & (1) \\ S(i+1, j) & (2) \\ S(i, j-1) & (3) \\ max_{i<k<j} S(i,k) + S(k+1, j) & (4) \end{cases}$$

$w(i,j) = 1$ if i, j are complementary (i.e., GC, CG, AU or UA); 0 otherwise

Dynamic programming

- Compute S(i,j) recursively (dynamic programming)
  - Compares a sequence against itself in a dynamic programming matrix
- Three steps

Initialization

	G	G	G	A	A	A	U	C	C
G	0								
G	0	0							
G	▨	0	0						
A	▨	▨	0	0					
A	▨	▨	▨	0	0				
A	▨	▨	▨	▨	0	0			
U	▨	▨	▨	▨	▨	0	0		
C	▨	▨	▨	▨	▨	▨	0	0	
C	▨	▨	▨	▨	▨	▨	▨	0	0

Example:

GGGAAAUCC

$S(i,i) = 0 \quad \forall \quad 1 \leq i \leq L$ → the main diagonal

$S(i,i-1) = 0 \quad \forall \quad 2 \leq i \leq L$ → the diagonal below

L: the length of input sequence

Recursion

Fill up the table (DP matrix) -- diagonal by diagonal

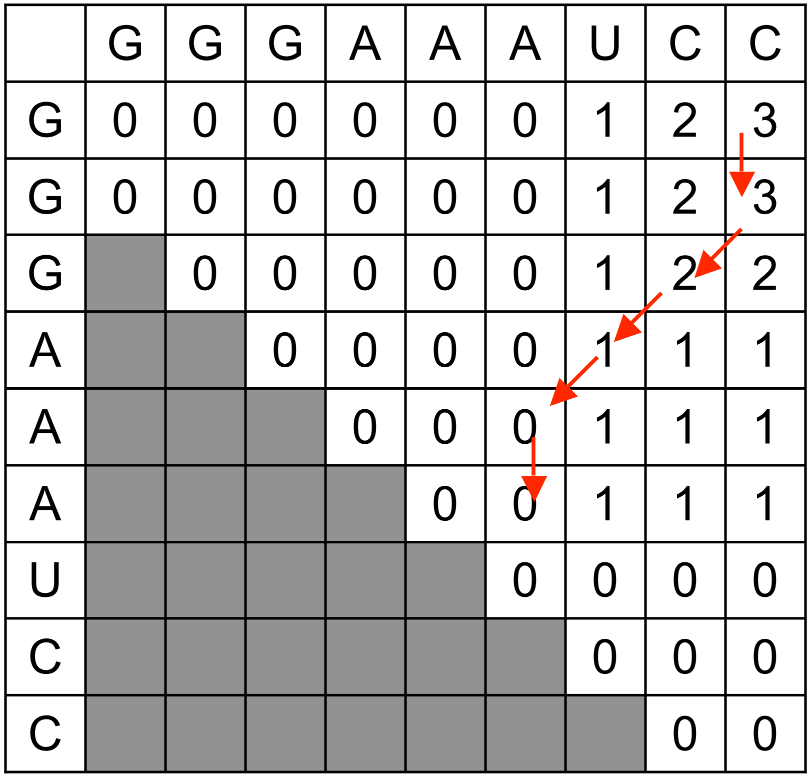
→ *j*

↓ *i*

	G	G	G	A	A	A	U	C	C
G	0	0	0	0					
G	0	0	0	0	0				
G		0	0	0	0	0			
A			0	0	0	0	?		
A				0	0	0	1		
A					0	0	1	1	
U						0	0	0	0
C							0	0	0
C								0	0

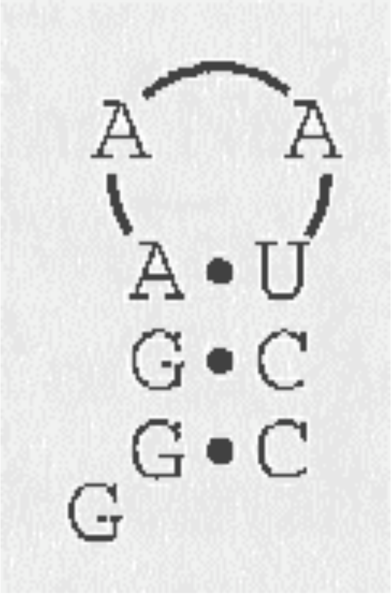
$$S(i,j) = max \begin{cases} S(i+1,j-1)+w(i,j) & (1) \\ S(i+1,j) & (2) \\ S(i,j-1) & (3) \\ max_{i<k<j} S(i,k)+S(k+1,j) & (4) \end{cases} \quad w(i,j) = \begin{cases} 1 & i,\ j \text{ are complementary} \\ 0 & otherwise \end{cases}$$

Traceback



	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

The structure is:



What are the other “optimal” structures?

Let’s play

Please bring in your sheet (with your inputs) to the class on Wed!!

- Input: AUGACAU
- Fill up the table
- Trace back

	A	U	G	A	C	A	U
A							
U							
G							
A							
C							
A							
U							

- Give the optimal structure
- What’s the size of the hairpin loop