In the Name of God, the Merciful, the Compassionate

Introduction to Bioinformatics
13 - RNA Structure Prediction

Instructor: Hossein Zeinali

Amirkabir University of Technology

Introduction

- RNA is one of the three major types of biological macromolecules.
- Understanding the structures of RNA provides insights into its functions.
  - Understanding the mechanisms of gene expression, viral infection, and immunity.
  - Is invaluable in drug design and understanding disease mechanisms.
- RNA structures can be experimentally determined using x-ray crystallography or NMR spectroscopy techniques.
  - Are extremely time consuming and expensive.
    - Computational prediction has become an attractive alternative.

RNA types & functions

- **RNA Functions:**
  - Storage/transfer of genetic information
  - Newly discovered regulatory functions - RNAi pathways
  - Catalytic

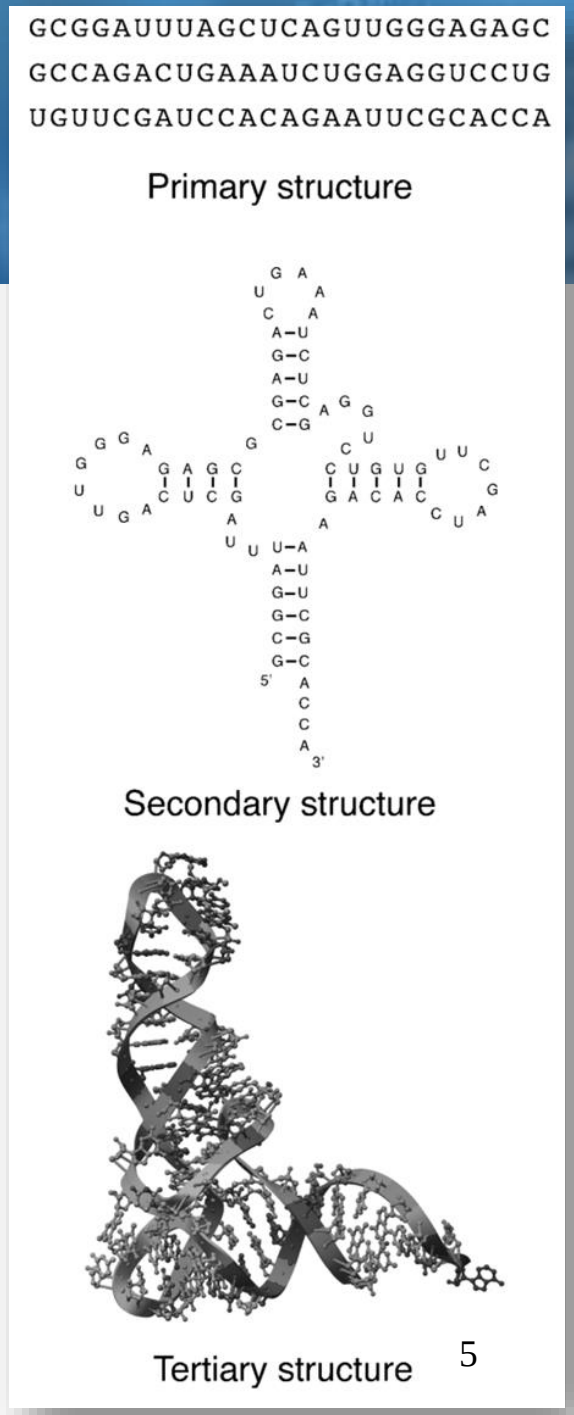
Types of RNAs	Primary Function(s)
mRNA - messenger	translation (protein synthesis) regulatory
rRNA - ribosomal	translation (protein synthesis) <catalytic>
tRNA - transfer	translation (protein synthesis)
hnRNA - heterogeneous nuclear	precursors & intermediates of mature mRNAs & other RNAs
scRNA – small cytoplasmic	signal recognition particle (SRP) tRNA processing <catalytic>
snRNA - small nuclear snoRNA - small nucleolar	mRNA processing, poly A addition <catalytic> rRNA processing/maturation/methylation
regulatory RNAs (siRNA, miRNA, etc.)	regulation of transcription and translation, other??

RNA Structure

- RNA is mainly single stranded
- The single RNA strand can self-hybridize to form base paired regions
- Generally, mRNA is more or less linear and non-structured, whereas rRNA and tRNA can only function by forming particular secondary and tertiary structures.
- Computational-based analysis is a main tool in RNA-based drug design in pharmaceutical industry.
- Knowledge of the secondary structures of rRNA is key for RNA-based phylogenetic analysis.

Levels of RNA Structure

- Like proteins, RNA has primary, secondary, and tertiary structures
  - Primary structure – linear base sequence of adenine (A), cytosine (C), guanine (G), and uracil (U).
  - Secondary structure - base-paired regions among single-stranded regions
    - Mainly composed of traditional Watson–Crick base pairing, which is A–U and G–C.
    - Plus non-canonical base pairing such as G and U (Wobble pair) base paring which is less stable.
  - Tertiary structure - 3D arrangement of bases of the RNA molecule.

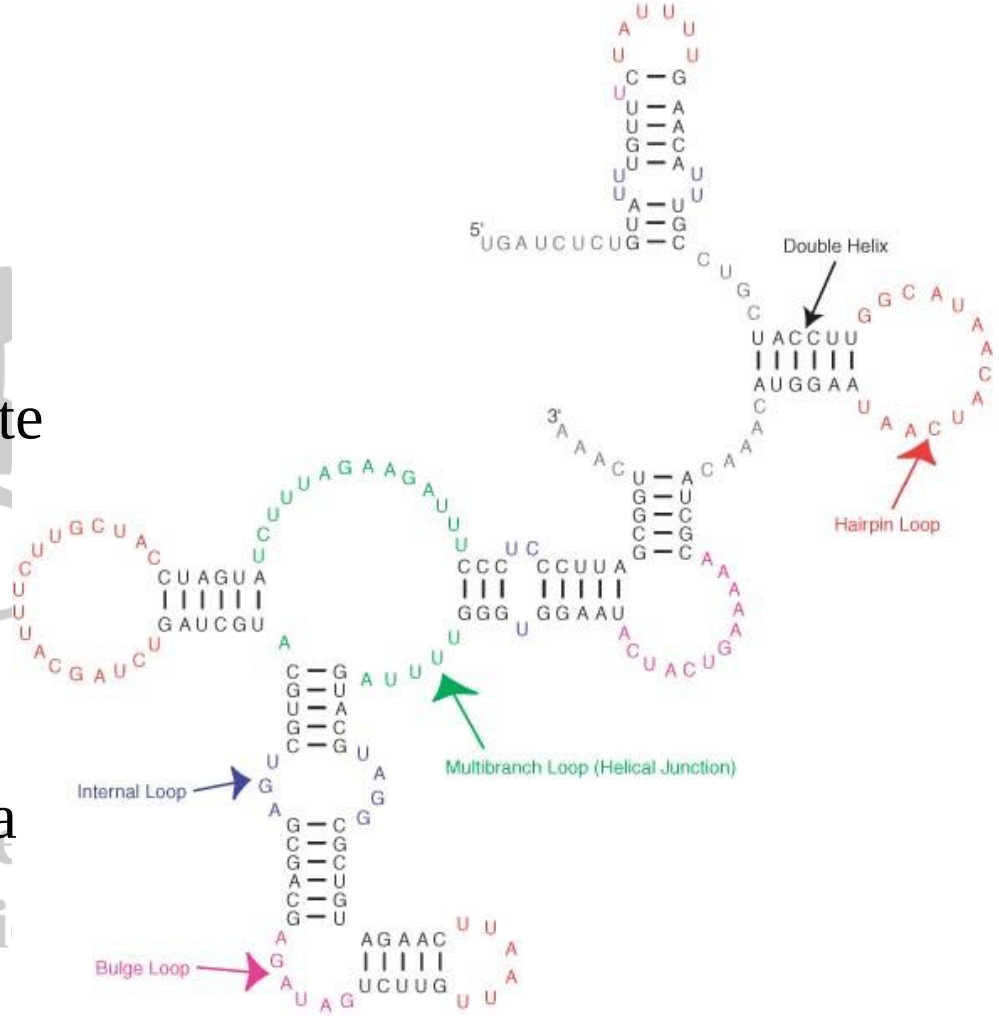


RNA Structure Prediction

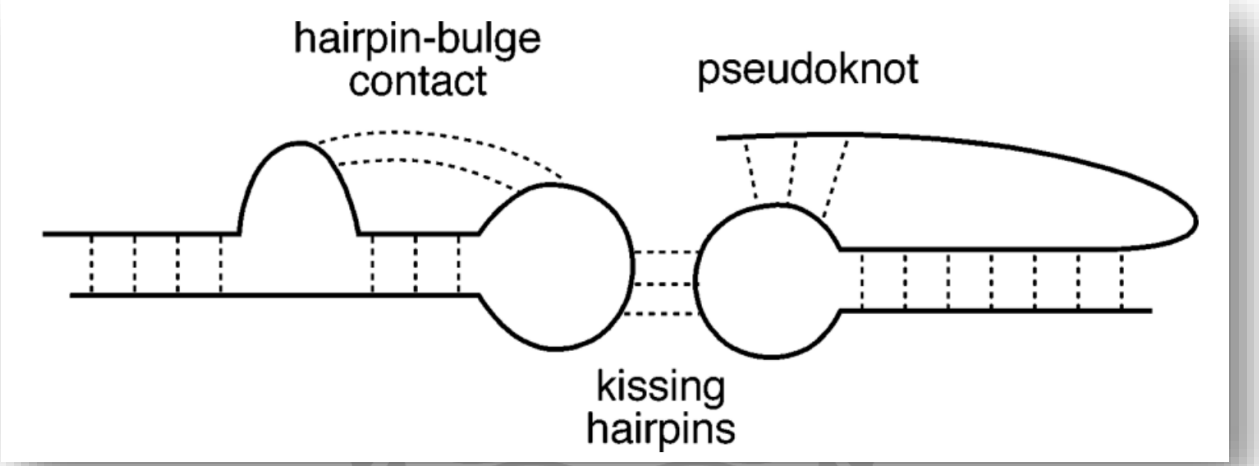
- RNA tertiary structure is very difficult to predict
  - Focus on predicting RNA secondary structure
- Given a RNA sequence, predict the secondary structure of the molecule
- Almost all methods ignore higher order secondary structures like pseudoknots
  - Are relatively rare in real structures

Common Structural Motifs in RNA

- **Helices**
- Loops
  - **Hairpin**: a structure with two ends of a single-stranded region (loop) connecting a base-paired region (stem)
  - **Interior**: two single stranded regions on opposite strands connecting two adjacent base-paired segments.
  - **Bulge**: a single stranded region connecting two adjacent base-paired segments
  - **Multibranch**: a loop that brings three or more base-paired segments in close vicinity forming a multi-furcated structure.

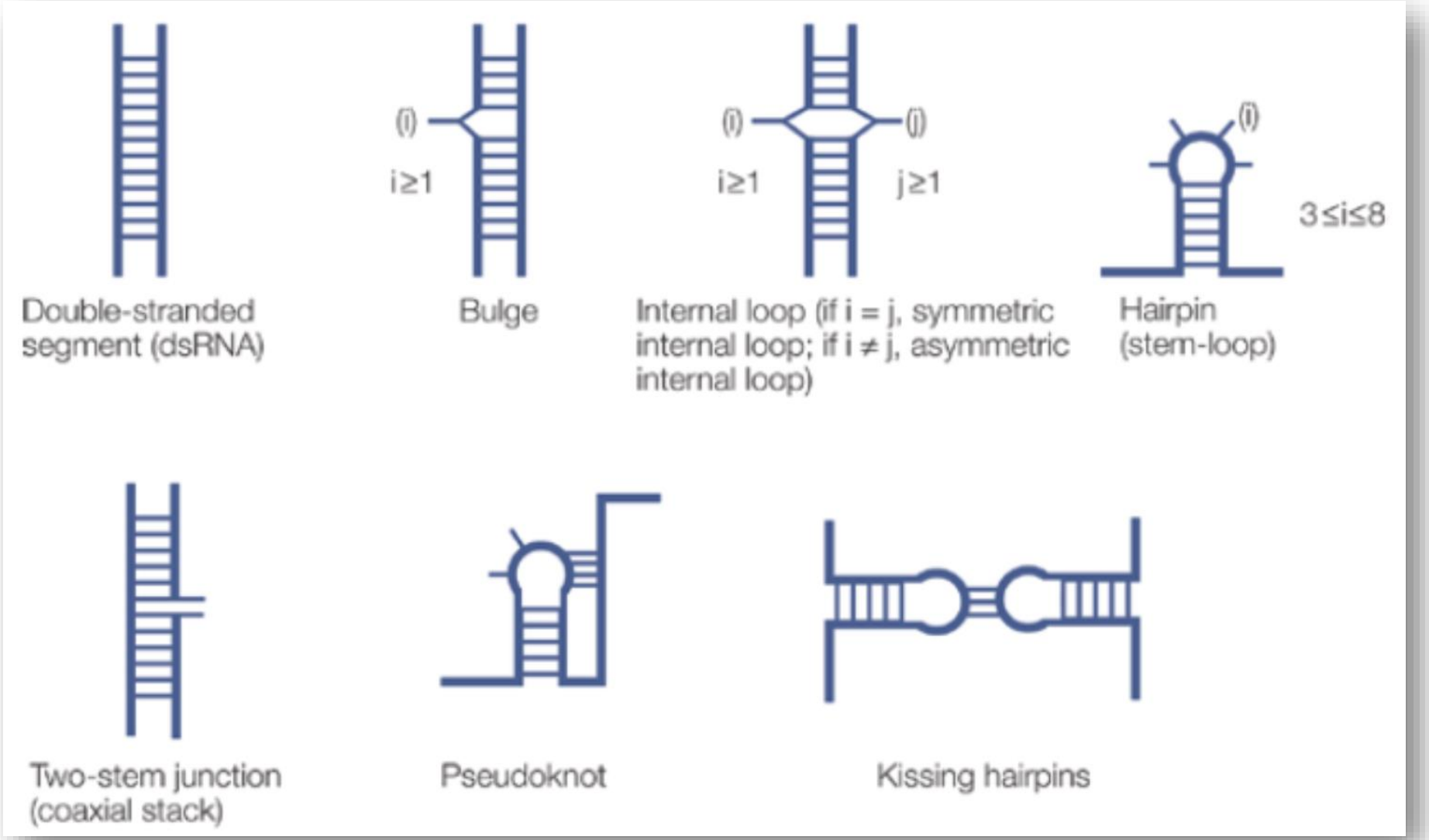


Higher Level of RNA Structures

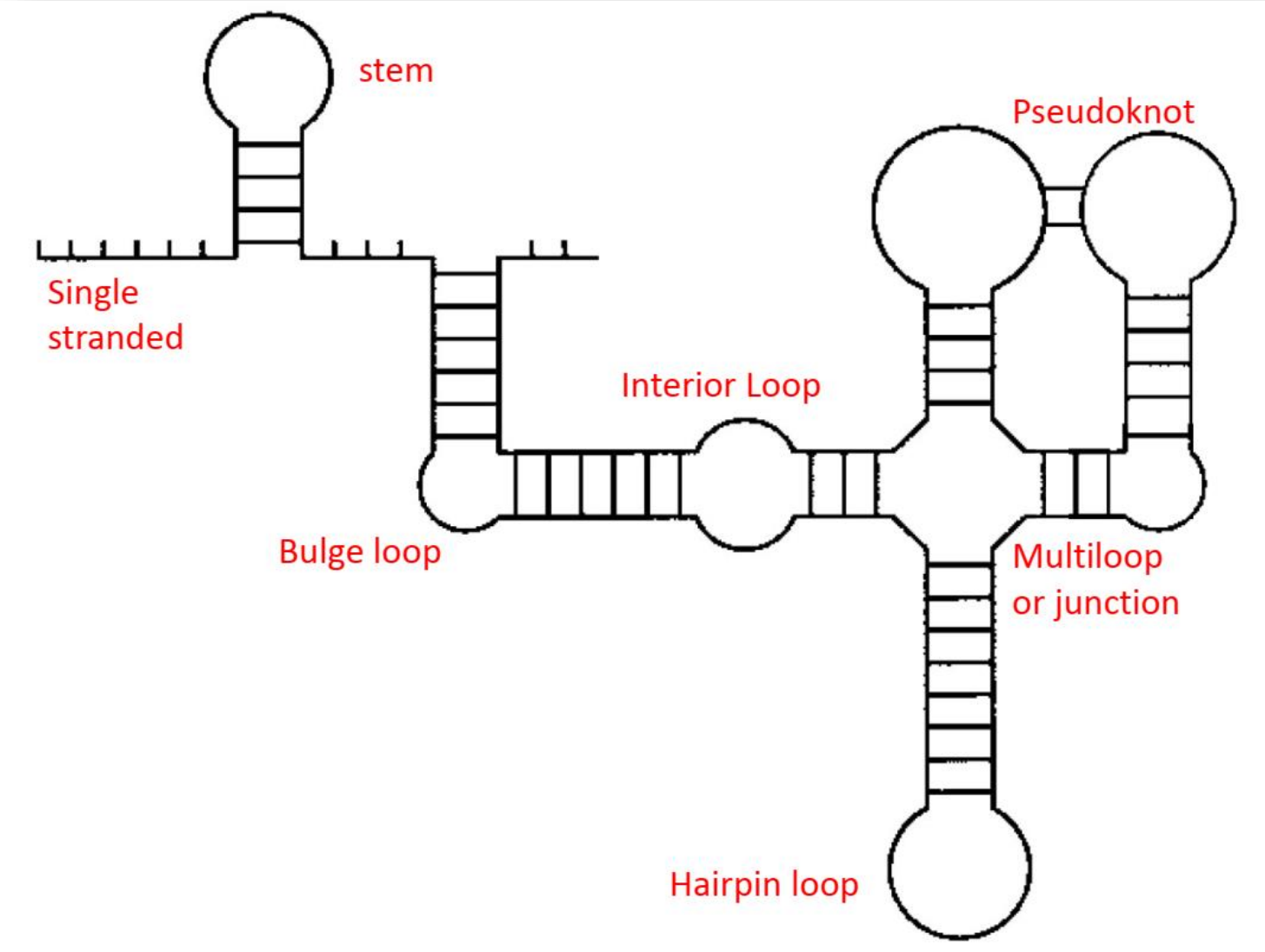


- Base pairing between loops of different secondary structural elements
  - **Pseudoknot loop**: base pairing formed between loop residues within a hairpin loop and residues outside the hairpin loop.
  - **Kissing hairpin**: a hydrogen bonded interaction formed between loop residues of two hairpin structures.
  - **Hairpin–bulge** contact: interactions between loop residues of a hairpin loop and a bulge loop.

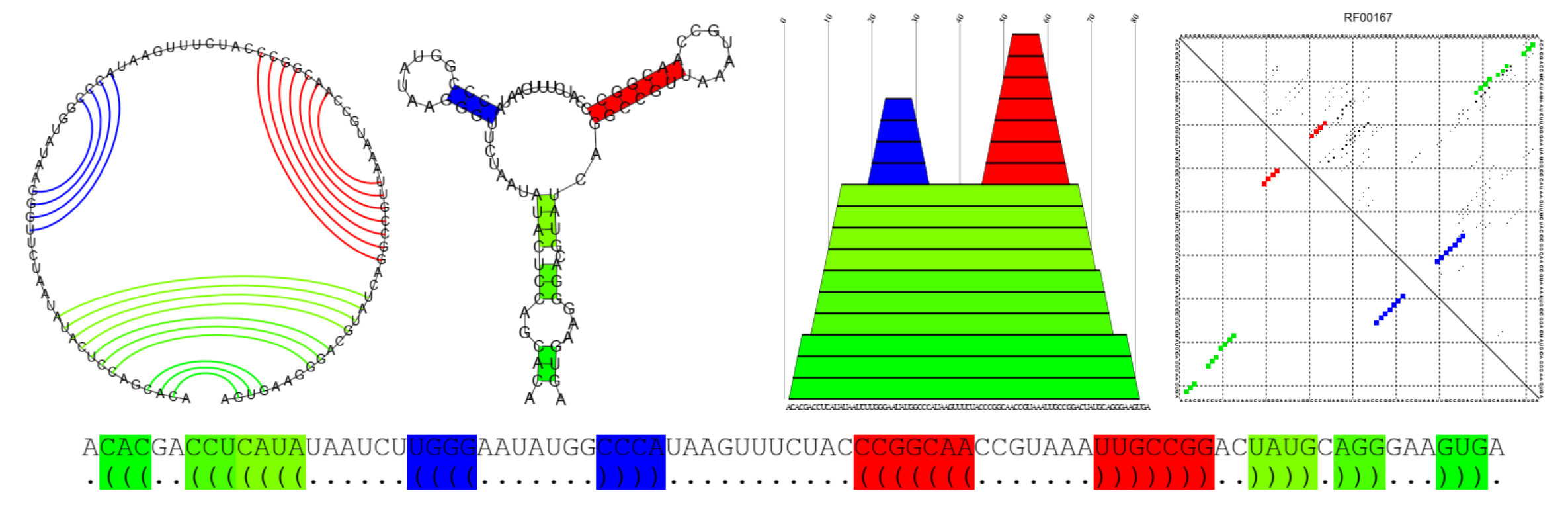
Secondary Structural Elements



RNA Structural Motifs



Representations of Secondary Structures



- From left to right: Circle plot, conventional secondary structure graph, mountain plot, dot plot.

RNA Secondary Structure Prediction Methods

- Three main types of methods:
- **Ab initio** - based on calculating most energetically favorable secondary structure(s)
  - Energy minimization (Zucker et al)
  - Maximize number of base pairs (Nussinov et al)
- **Comparative approach** - based on comparisons of multiple evolutionarily-related RNA sequences
  - Sequence comparison (covariation)
- **Combined** computational & experimental
  - Use experimental constraints when available

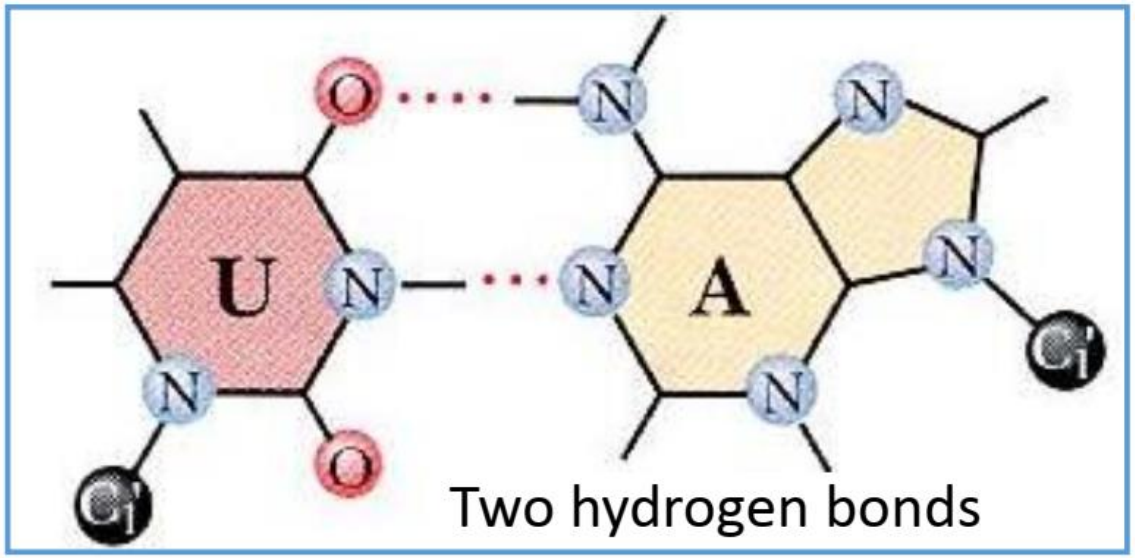
Ab Initio Prediction

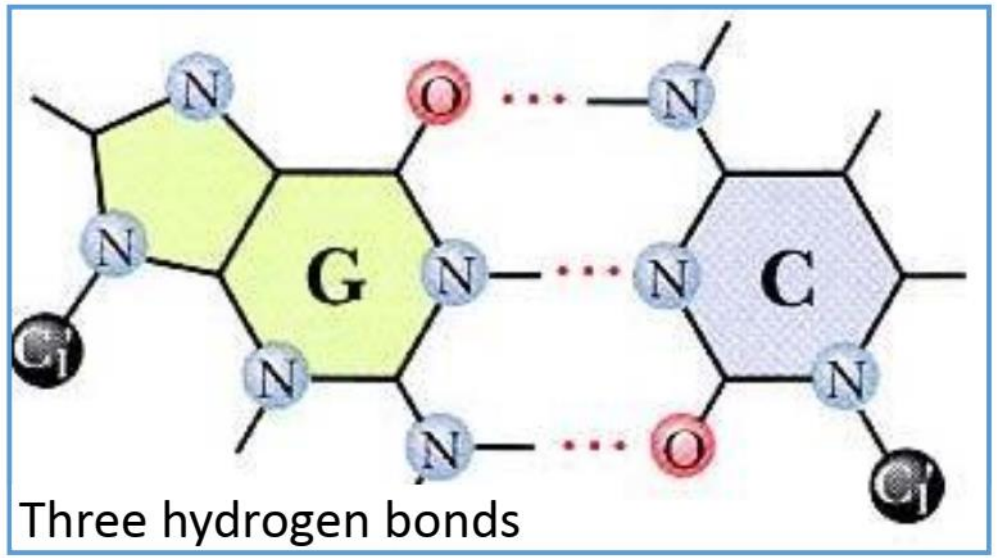
- Requires only a single RNA sequence
  - The structure of an RNA molecule is solely determined by its sequence.
- Finds a structure with minimum free energy
  - Generally, when a base pairing is formed, the energy of the molecule is lowered because of attractive interactions between the two strands.
- Base-paired regions have lower free energy, so methods "attempt to find secondary structure with maximal base pairing”
  - Note: largest contribution to energy is nearest neighbor (base-stacking) interactions, not base-pairing!

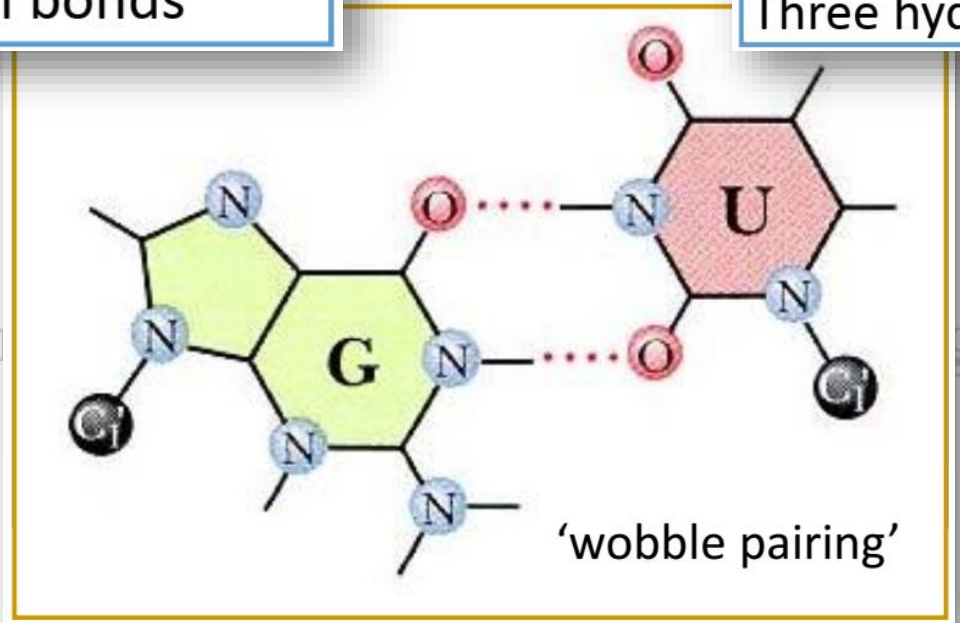
Ab Initio Prediction

- Free energy can be calculated based on parameters empirically derived for small molecules.
  - G–C base pairs (~3 kcal/mol) are more stable than A–U base pairs (~2 kcal/mol), which are more stable than G–U base pairs (~1 kcal/mol).
- Base-pair formation is not independent: multiple base-pairs adjacent to each other are more favorable than individual base-pairs - ***cooperative - because of base-stacking interactions***
- **Bulges** and **loops** adjacent to base-pairs have a free energy ***penalty***
  - The neighboring loops and bulges tend to destabilize the base-pair formation.

Canonical Base Pairs





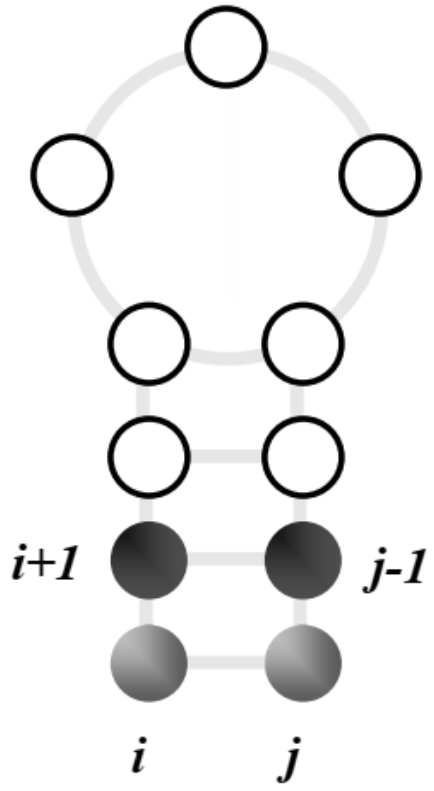


Dynamic Programming (Nussinov Algorithm)

- Finding optimal secondary structure is difficult - lots of possibilities
  - Dynamic programming can be used for this aim
- Compare RNA sequence with itself
- Construct the structure step by step and add new bases to the best-found sub-structure at each step. We will see 4 types for adding bases.
- Apply scoring scheme based on energy parameters for base stacking, cooperativity, and penalties for destabilizing forces
- Find path that represents most energetically favorable secondary structure

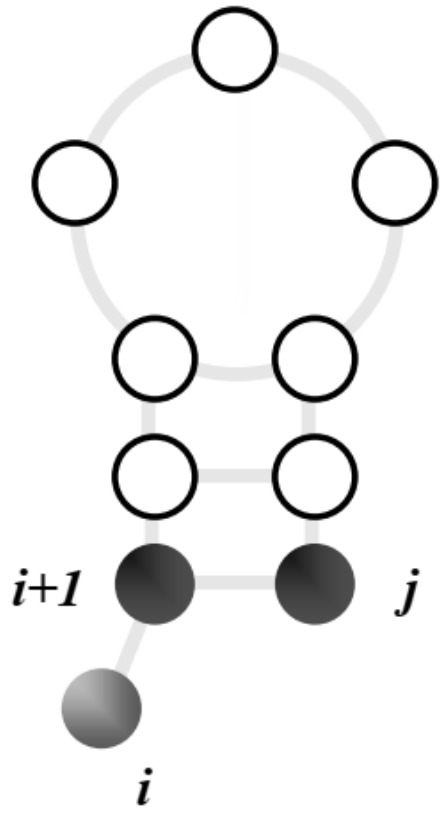
Dynamic Programming: Type1

- Adding base pare (i, j) to the best found sub-structure for subsequence (i+1, j-1)



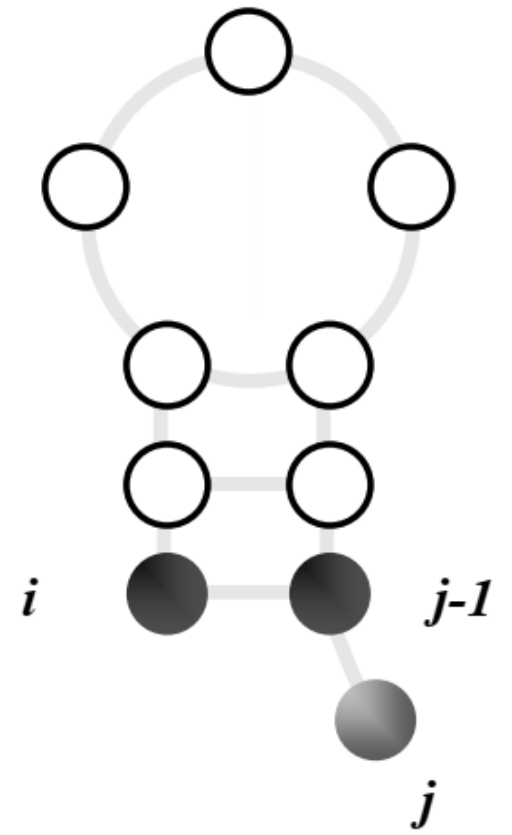
Dynamic Programming: Type2

- Adding a single base at position i to the best found sub-structure for subsequence (i+1, j)



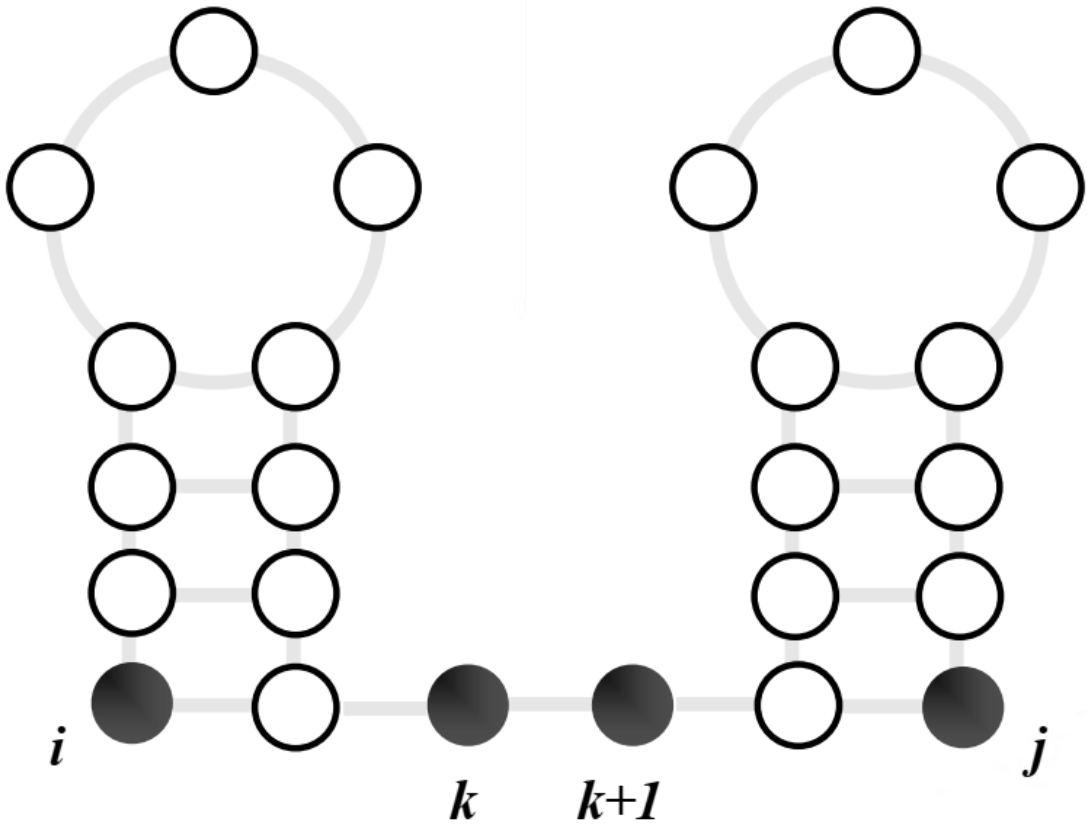
Dynamic Programming: Type3

- Adding a single base at position j to the best found sub-structure for subsequence (i, j-1)



Dynamic Programming: Type4

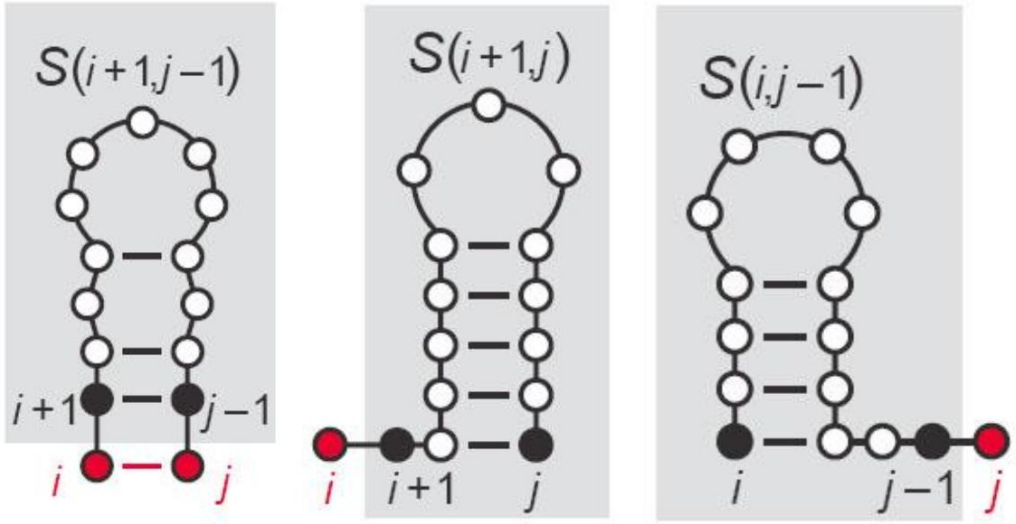
- Combining 2 optimal sub-structures for subsequences (i, k) and (k+1, j)



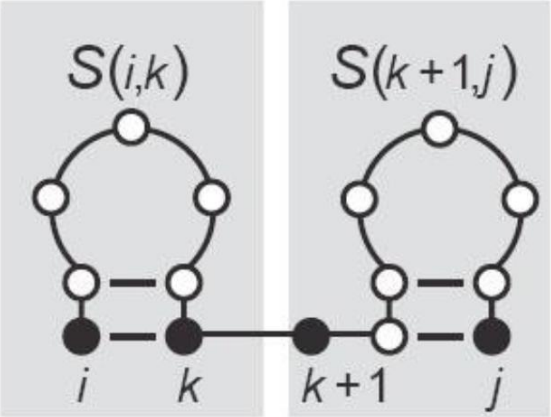
Dynamic Programming

- *S(i, j)* is the folding of the RNA subsequence of the strand from index *i* to index *j* which results in the highest number of base pairs.
- Recurrence:

$$S(i,j) = \max \begin{cases} S(i+1, j-1) + 1 \\ S(i+1, j) \\ S(i, j-1) \\ \max\limits_{i<k<j} S(i,k) + S(k+1, j) \end{cases}$$



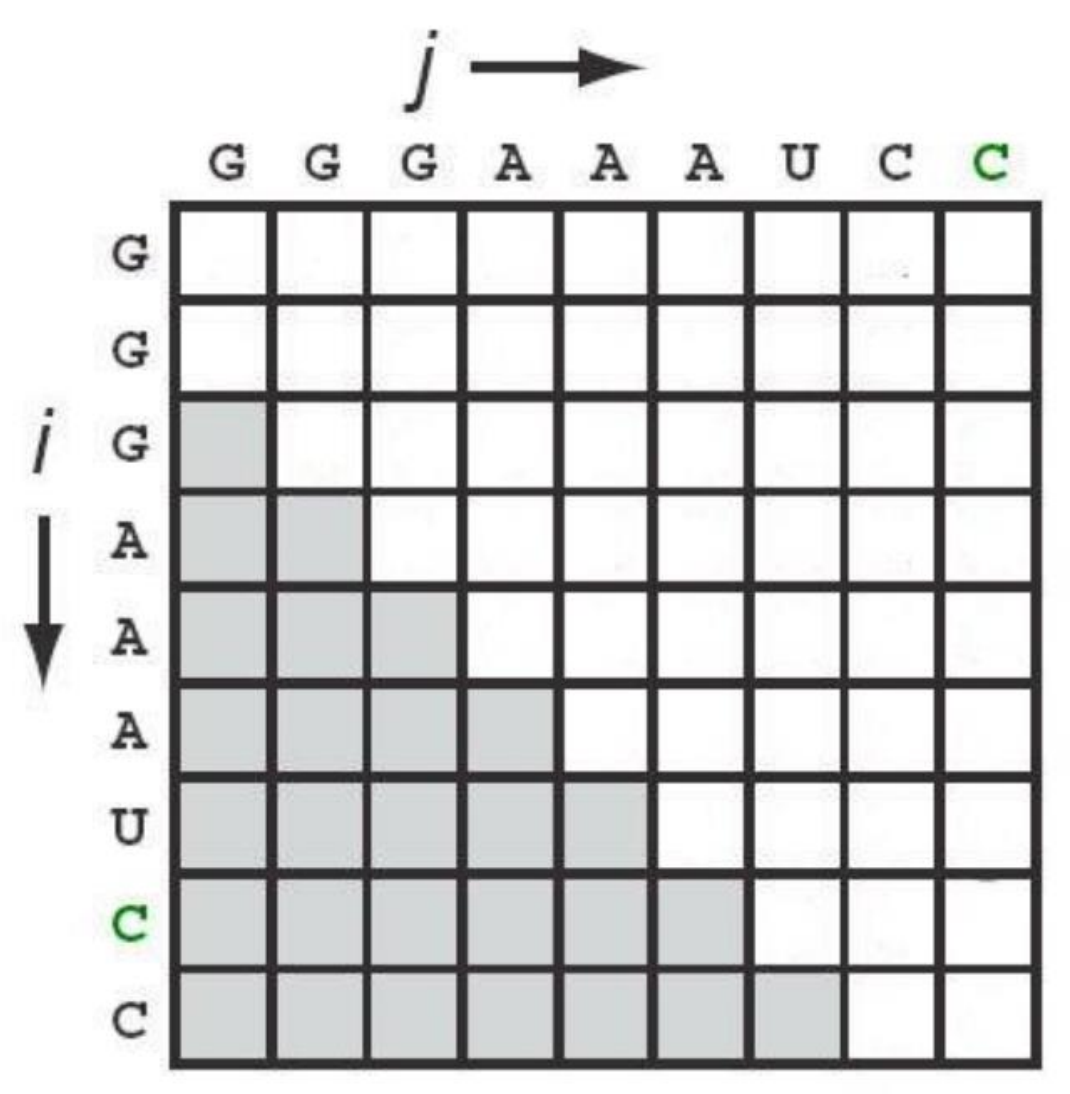
Base pair at *i* and *j* | Unmatched at *i* | Unmatched at *j*



Bifurcation

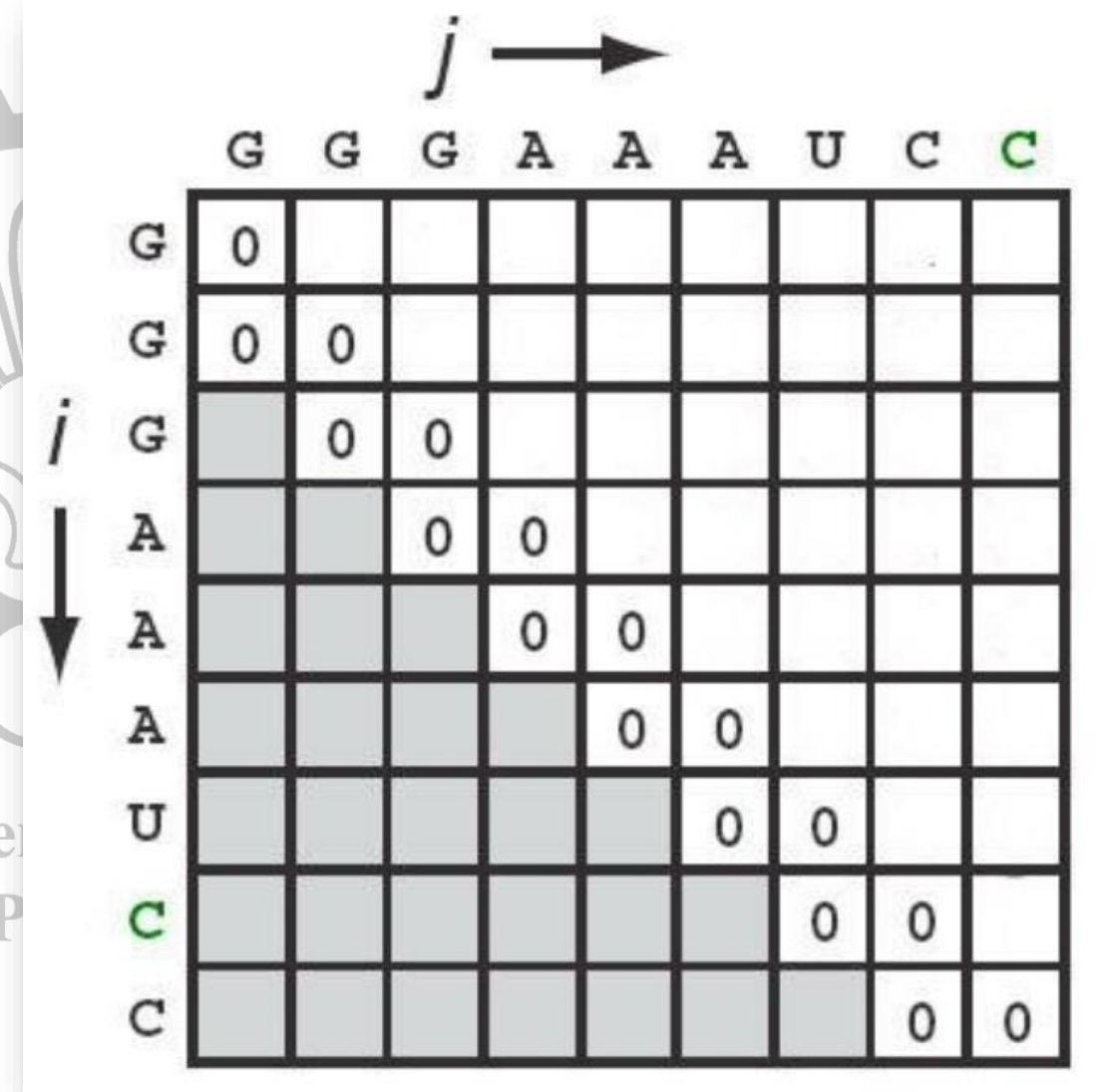
Example

- Alignment Method:
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score is independent of overall structure



Example

- Initialize first two diagonals to 0



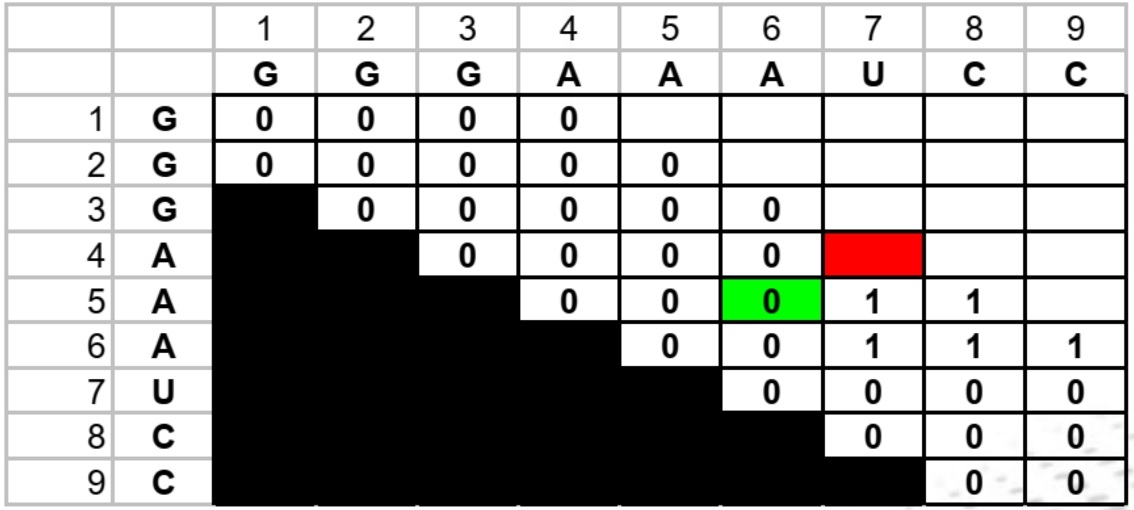
Example

- Fill in squares sweeping diagonally

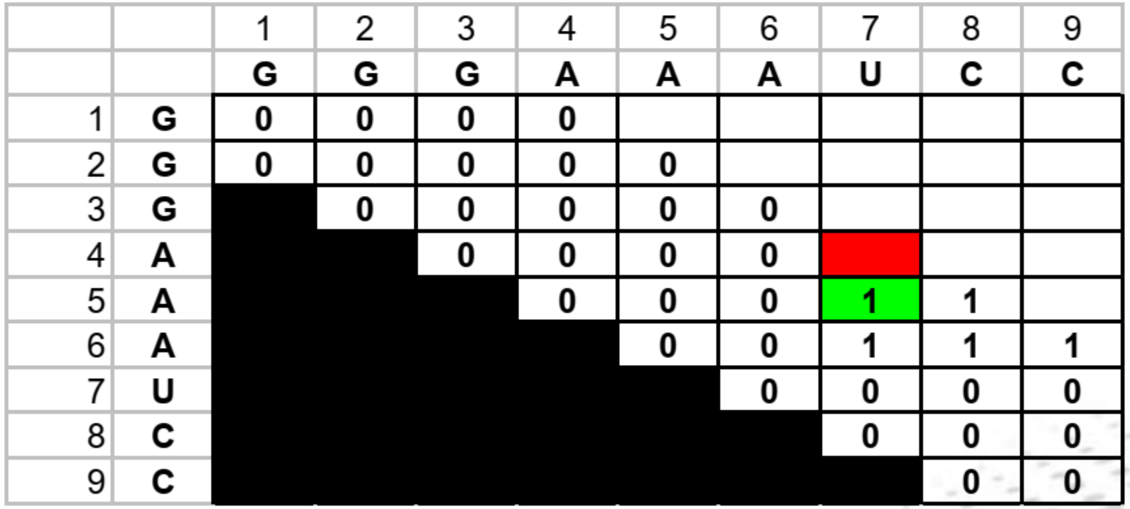
$$S(i,j) = \max \begin{cases} S(i+1, j-1) + 1 \\ S(i+1, j) \\ S(i, j-1) \\ \max\limits_{i<k<j} S(i,k) + S(k+1, j) \end{cases}$$

i \ j	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	0
C								0	0

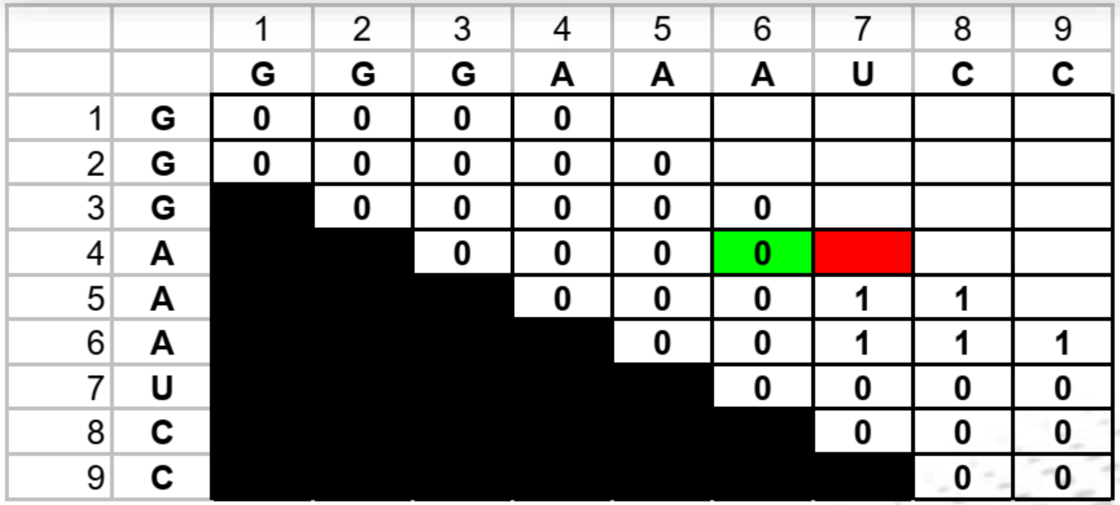
Comparison of 4 Types



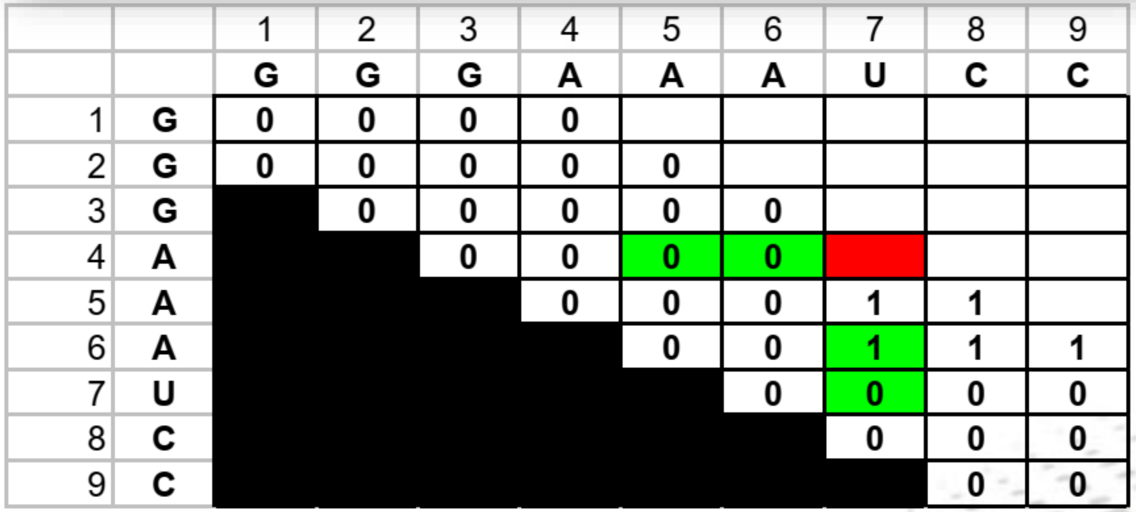
		1	2	3	4	5	6	7	8	9
		G	G	G	A	A	A	U	C	C
1	G	0	0	0	0					
2	G	0	0	0	0	0				
3	G		0	0	0	0	0			
4	A			0	0	0	0	🟥		
5	A				0	0	0	1	1	
6	A					0	0	1	1	1
7	U						0	0	0	0
8	C							0	0	0
9	C								0	0



		1	2	3	4	5	6	7	8	9
		G	G	G	A	A	A	U	C	C
1	G	0	0	0	0					
2	G	0	0	0	0	0				
3	G		0	0	0	0	0			
4	A			0	0	0	0	🟥		
5	A				0	0	0	1	1	
6	A					0	0	1	1	1
7	U						0	0	0	0
8	C							0	0	0
9	C								0	0



		1	2	3	4	5	6	7	8	9
		G	G	G	A	A	A	U	C	C
1	G	0	0	0	0					
2	G	0	0	0	0	0				
3	G		0	0	0	0	0			
4	A			0	0	0	0	🟥		
5	A				0	0	0	1	1	
6	A					0	0	1	1	1
7	U						0	0	0	0
8	C							0	0	0
9	C								0	0



		1	2	3	4	5	6	7	8	9
		G	G	G	A	A	A	U	C	C
1	G	0	0	0	0					
2	G	0	0	0	0	0				
3	G		0	0	0	0	0			
4	A			0	0	0	0	🟥		
5	A				0	0	0	1	1	
6	A					0	0	1	1	1
7	U						0	0	0	0
8	C							0	0	0
9	C								0	0

Example

- Fill in squares sweeping diagonally

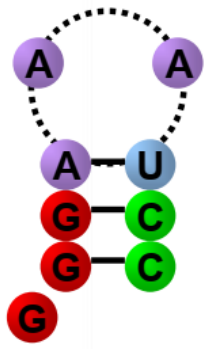
$$S(i,j) = \max \begin{cases} S(i+1, j-1) + 1 \\ S(i+1, j) \\ S(i, j-1) \\ \max\limits_{i<k<j} S(i,k) + S(k+1, j) \end{cases}$$

		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
i	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

$j \rightarrow$

Traceback

- Start from (1, L) and backtrack to hit diagonal



$j \longrightarrow$

$i \downarrow$

		1	2	3	4	5	6	7	8	9
		G	G	G	A	A	A	U	C	C
1	G	0	0	0	0	0	0	1	2	3
2	G	0	0	0	0	0	0	1	2	3
3	G		0	0	0	0	0	1	2	2
4	A			0	0	0	0	1	1	1
5	A				0	0	0	1	1	1
6	A					0	0	1	1	1
7	U						0	0	0	0
8	C							0	0	0
9	C								0	0

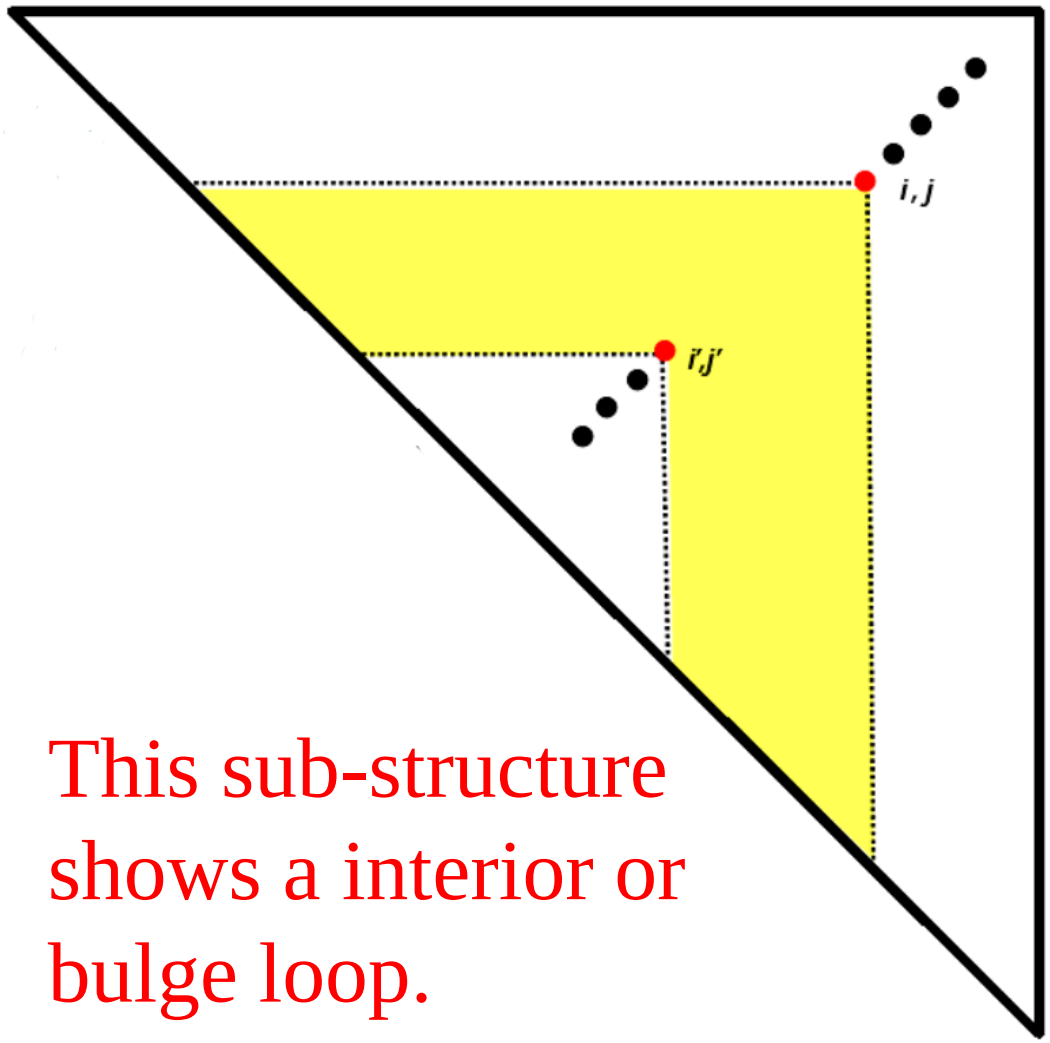
Problem with DP Approach

- DP returns SINGLE lowest energy structure
- There may be many structures with near minimum energy but not necessarily the one with maximum base pairs.
- Also, *predicted* secondary structure is only as good as ***energy parameters*** used
- ***Solution:*** return multiple structures with near optimal energies
- Base pair maximization will not necessarily lead to the most stable structure.
  - It may create structure with many interior loops or hairpins which are energetically unfavorable

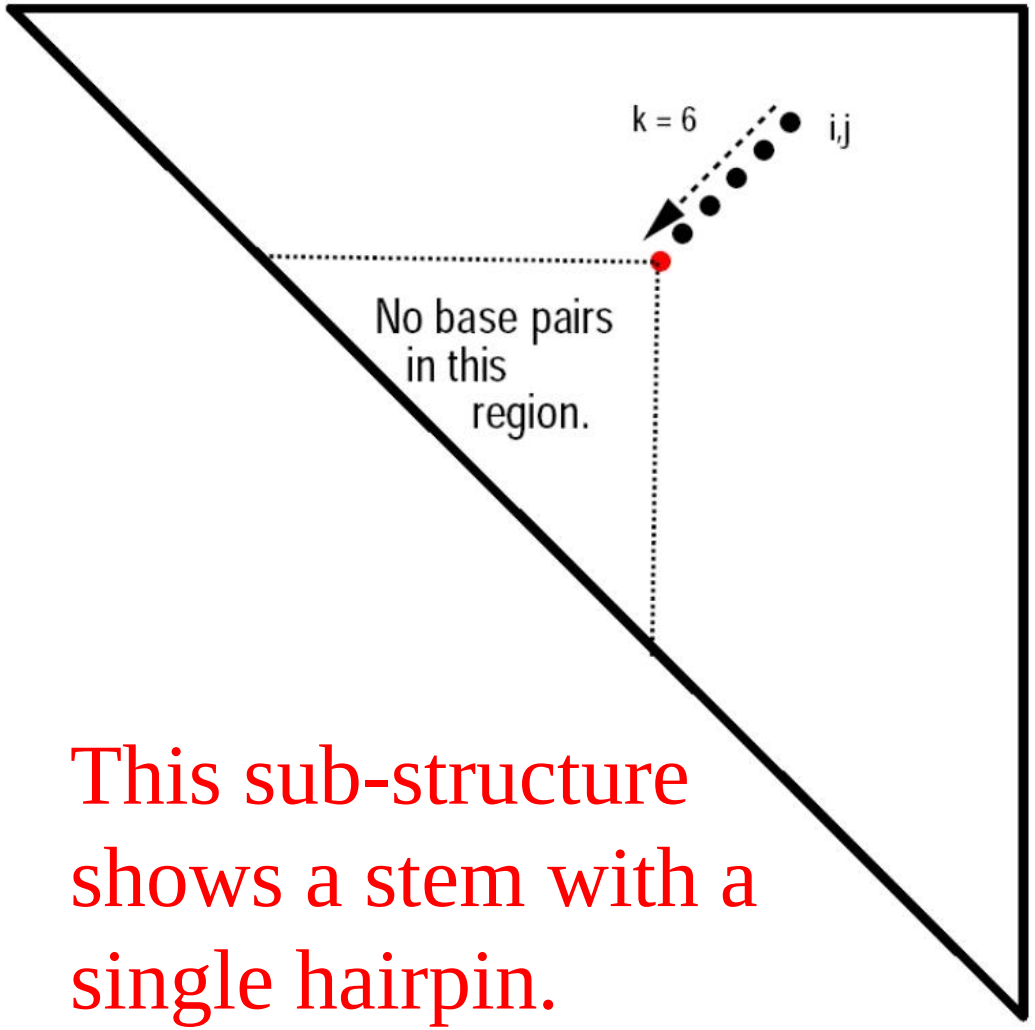
Dot Matrices

- In searching for the lowest energy form, all possible base-pair patterns have to be examined.
  - The dot matrix method and the dynamic programming method can be used
- Compare input sequence to itself and put a dot where there is a complimentary base
- The diagonals perpendicular to the main diagonal represent regions that can self-hybridize to form double-stranded structure.
- Is often obscured by high noise levels.
  - Can be reduced by windowing

Sub-structures in Dot-Matrix

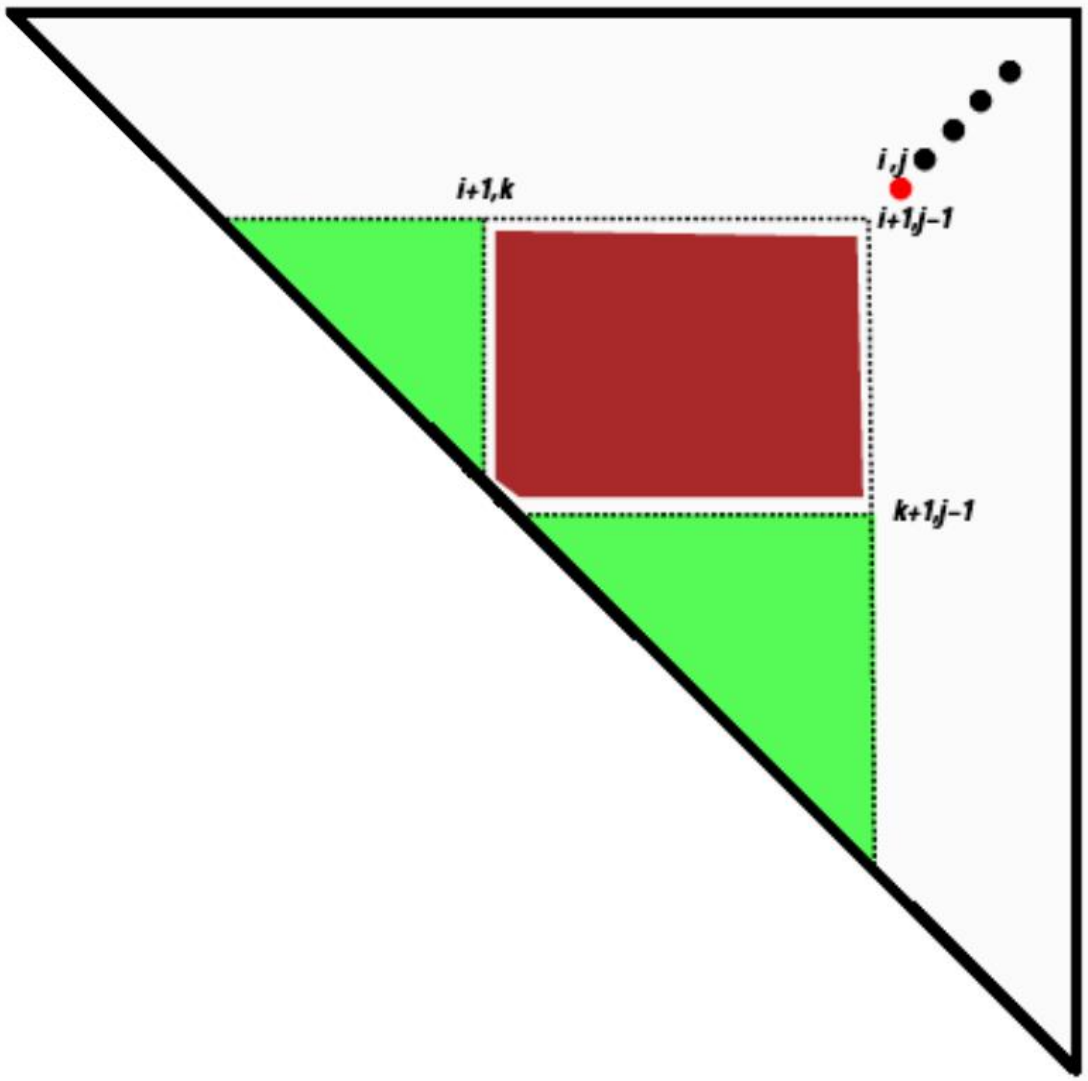


This sub-structure shows a interior or bulge loop.



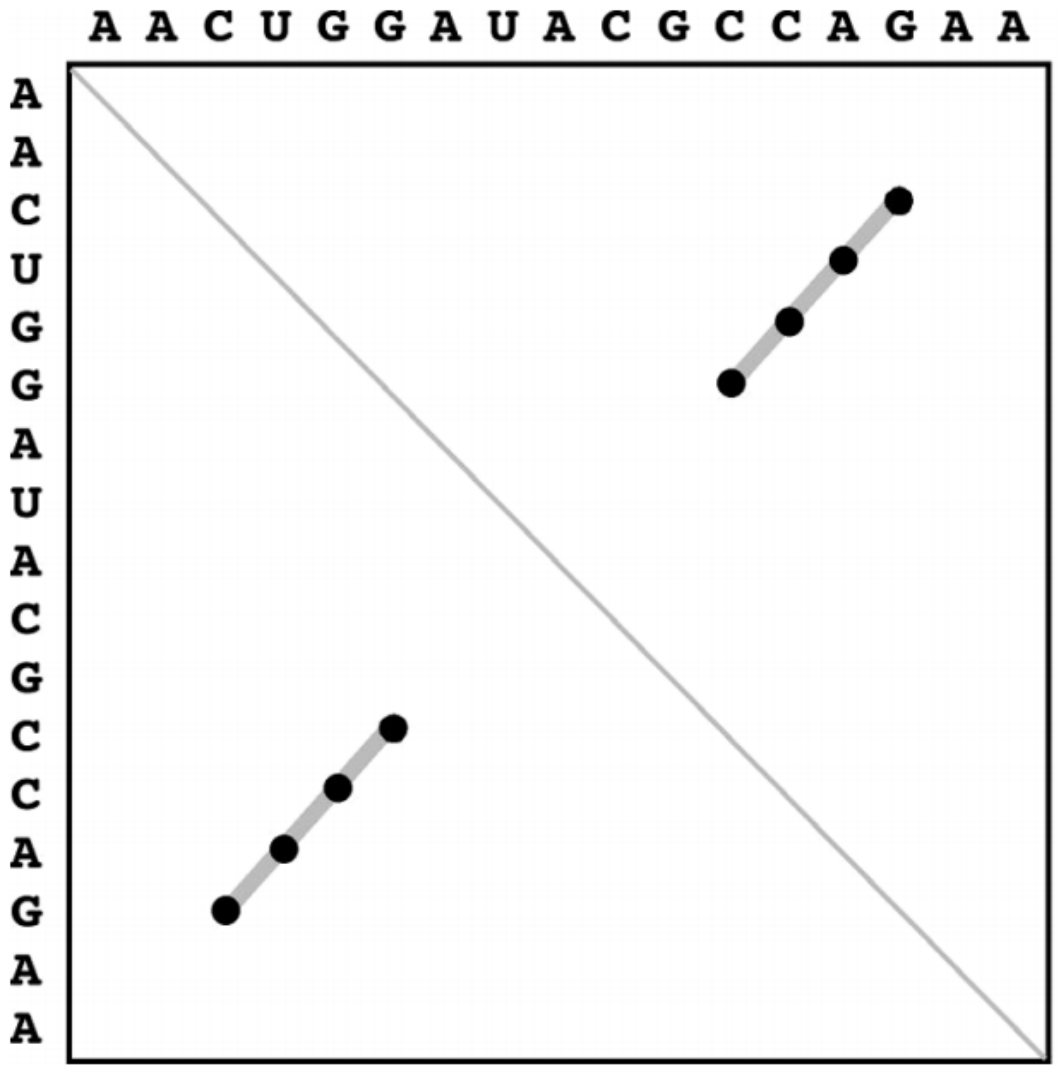
This sub-structure shows a stem with a single hairpin.

Sub-structures in Dot-Matrix



- This sub-structure shows a multi loop
  - There is no any pairs in the brown area.
  - There should be some pairs in the green areas.

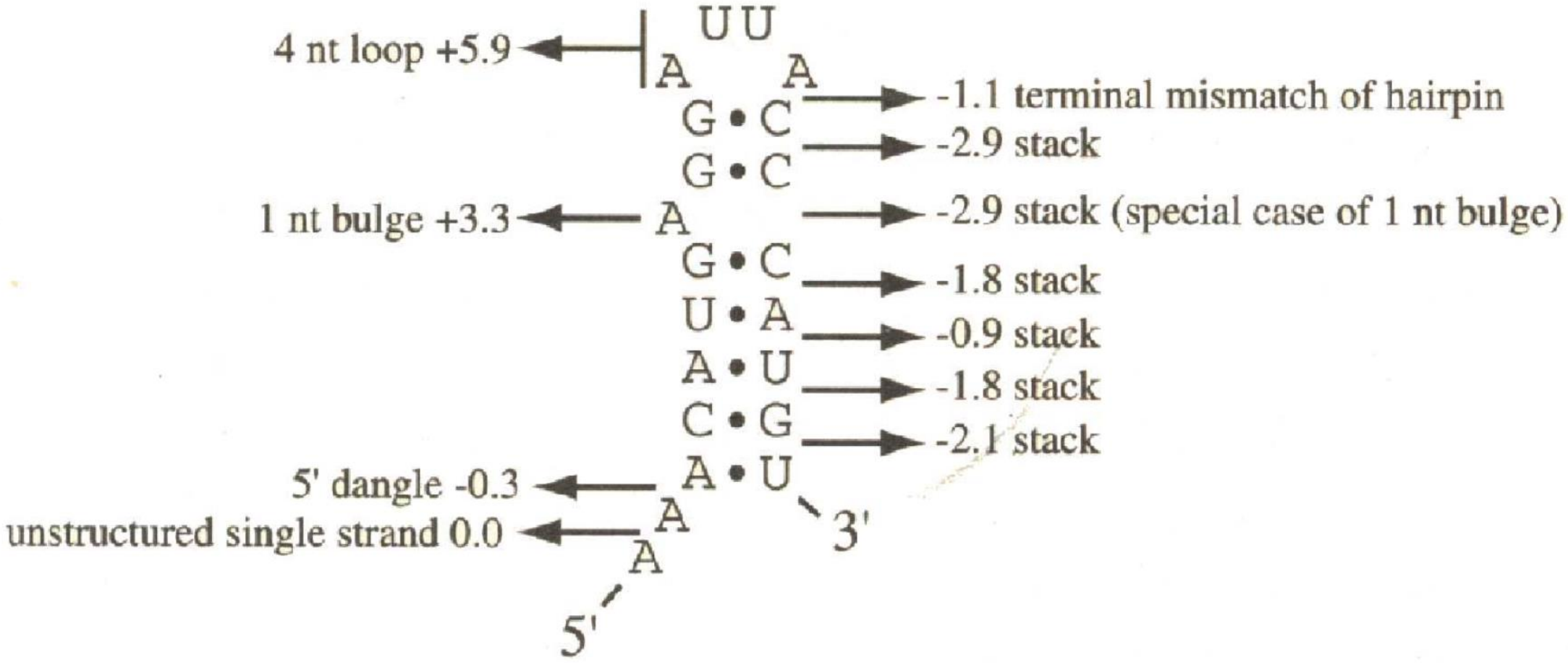
Dot Matrix Example



Minimum Free Energy: Zuker Algorithm

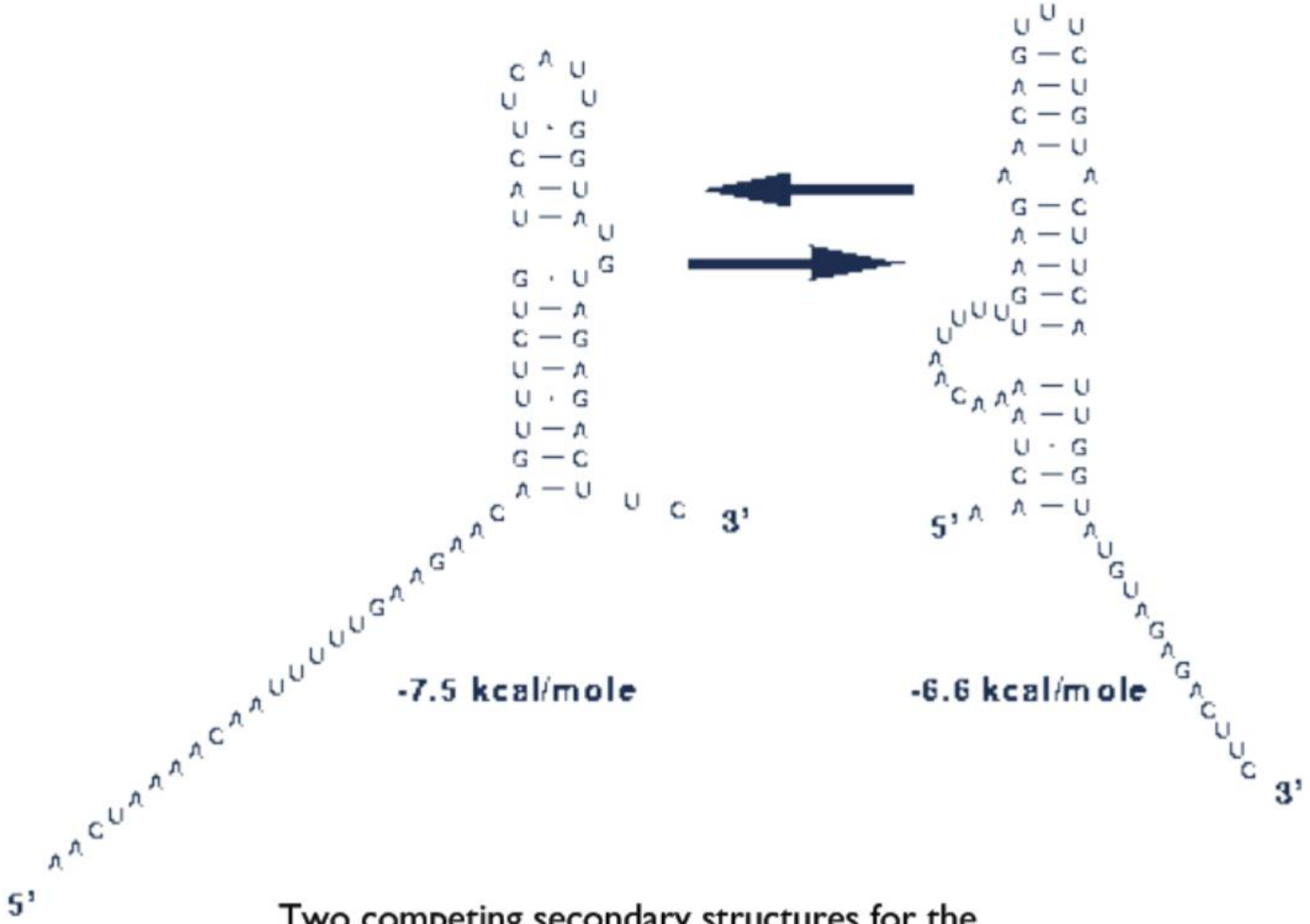
- Overcome the main drawback of Nussinov's algorithm: non-realism of base pair maximization!
- Define an energy model for RNA that can be parameterized by experimentally measured energies
- Devise an algorithm that minimizes the free energy of RNA according to this model
- The algorithm is using dynamic programming in a similar way to Nussinov's algorithm

Free Energy: an Example



overall ΔG = -4.6 kcal/mol

RNA Structure Comparison



Two competing secondary structures for the Leptomonas collosoma spliced leader mRNA.

Popular *Ab Initio* Prediction Programs

- **Mfold**
  - www.bioinfo.rpi.edu/applications/mfold/
  - Combines DP with thermodynamic calculations
  - It also produces dot plots coupled with energy terms.
  - Fairly accurate for short sequences, less accurate as sequence length increases

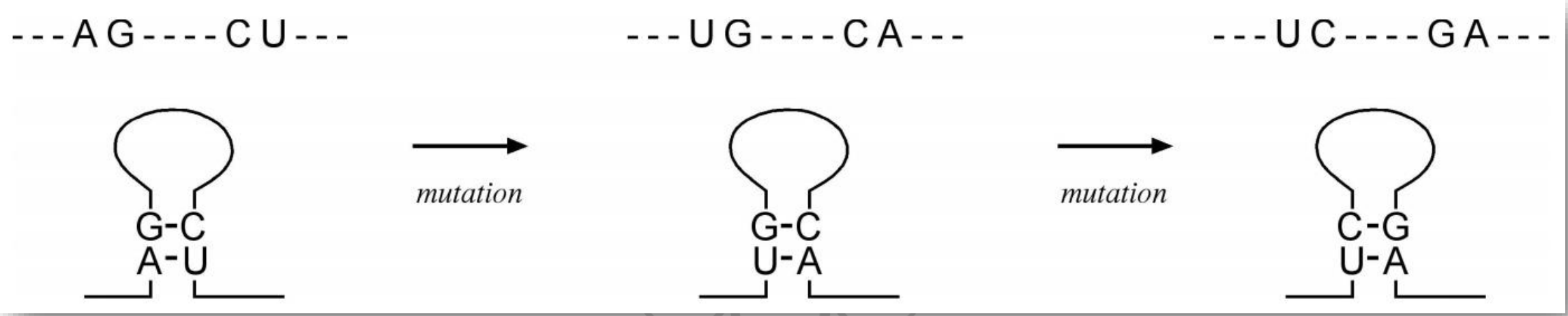
- **RNAfold**
  - http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi
  - Returns multiple structures near predicted optimal structure
  - Computes larger number of potential secondary structures than Mfold, so uses a simplified energy function
    - The prediction results are not always guaranteed to be better than those predicted by Mfold.

Comparative Prediction Approaches

- Use multiple sequence alignment
- Assume related RNA sequences (homologous) fold into same secondary structure
- RNA functional motifs are structurally conserved
- ***Covariation*** concept: to maintain RNA structure during evolution, a mutation in a base-paired residue must be **compensated** for by a mutation in residue with which it pairs
- Comparative methods search for covariation patterns in MSAs
- Predict secondary structure of each individual sequence in a MSA
- Compare all structures and try to identify a consensus structure

Comparative Prediction Approaches

- Use multiple sequence alignment
- Assume related RNA sequences (homologous) fold into same



- Comparative methods search for covariation patterns in MSAs
- Predict secondary structure of each individual sequence in a MSA
- Compare all structures and try to identify a consensus structure

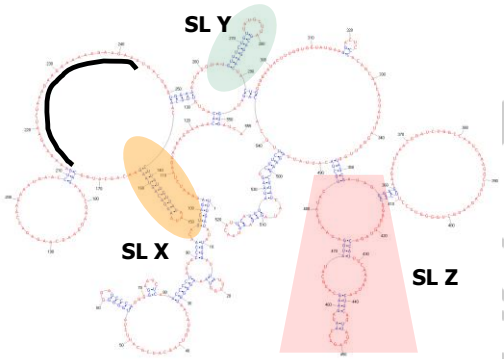
Popular Comparative Prediction Programs

- **RNAalifold**:
  - http://rna.tbi.univie.ac.at/cgi-bin/alifold.cgi
  - Requires user to provide MSA as a *pre-alignment*
  - Creates a scoring matrix combining minimum free energy and covariation information
  - DP used to identify minimum free energy structure
  - Is relatively successful for reasonably conserved sequences and depends on the quality of the input alignment.

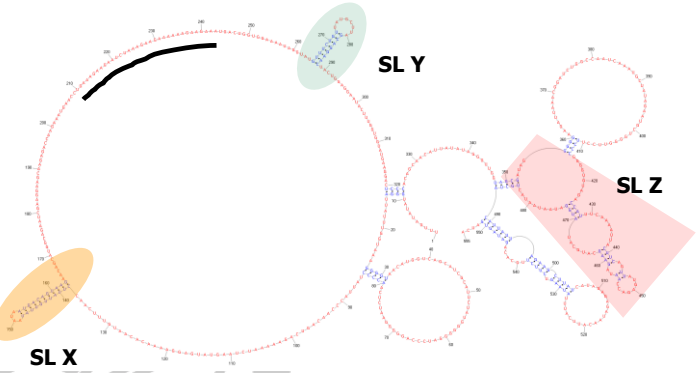
Popular Comparative Prediction Programs

- **Foldalign**:
  - http://foldalign.kvl.dk/server/index.html
  - User provides pair of unaligned RNA sequences
    - Two sequences limitation is due to dynamic programming computation
  - Constructs alignment using dynamic programming & computes conserved structure
  - Suitable only for relatively short sequences
- **Dynalign**:
  - http://rna.urmc.rochester.edu
  - Does not require sequence similarity and therefore can handle very divergent sequences.

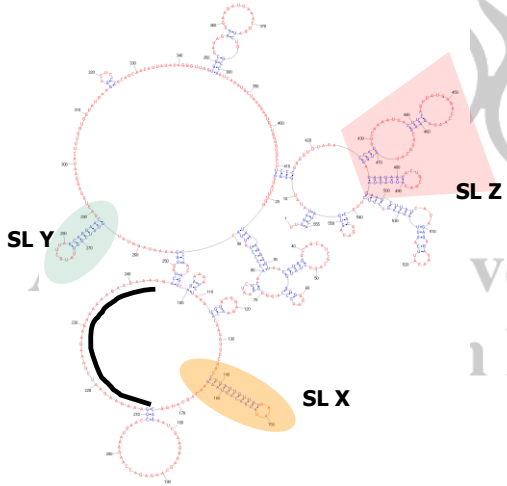
Comparison of Predictions using Different Methods



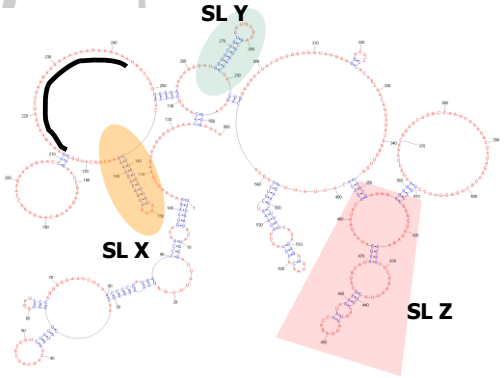
Mfold -54.84 kcal/mol



Sfold -51.14 kcal/mol



RNAstructure -71.3 kcal/mol



RNAfold -80.16 kcal/mol

Performance Evaluation

- **Correlation coefficient**: takes into account both sensitivity and selectivity information
- Ab initio methods? correlation coefficient = 20-60% depending on the length
- Comparative approaches? correlation coefficient = 20-80%
- Programs that require user to supply MSA are more accurate
- Comparative programs are consistently more accurate than ab initio
  - For small RNA sequences such as tRNA, both subtypes can achieve very high accuracy (up to 100%).

- BEST APPROACH? Methods that combine computational prediction (ab initio & comparative) with experimental constraints (from chemical/enzymatic modification studies)

References

- Mostly used:
  - Essential bioinformatics, Chapter 16 (RNA Structure Prediction)

- IP notice: some slides were selected from Drena Dobbs' and Ziv Bar-Joseph's slides.

Thanks for your attention