Bioinformatics Algorithms
COS-BIOL-530/630
Lecture09

Days & Times	Room	Meeting Dates
Tu 2:00PM - 3:50PM	Thomas Gosnell Hall (GOS)-2178	01/13/2025 - 04/28/2025
Th 2:00PM - 3:50PM	Thomas Gosnell Hall (GOS)-2178	01/13/2025 - 04/28/2025

Instructor:
Fernando Rodriguez
email: frvsbi@rit.edu
Office: Orange Hall 1311

RNA secondary structure
- Lecture09-

Announcements

Lecture09
Lab09
- Activity 9
- Discussion 9

Quiz 8 (Lecture/Lab09) opens Friday, March 21st

Thursday Lab09

!

- **JupyterLab / Notebook: Lab09_Nussinov.ipynb (Content>Labs)**

https://jupyter.org/
https://jupyter.org/try-jupyter/lab/
https://code.visualstudio.com/

- ***RNAMotif***:

https://github.com/dacase/rnamotif

In *Oedipus*:
/mnt/sde_dir/software/rnamotif/rnamotif

RNA secondary structure - Lecture09-

Topics:

- RNA: structural and catalytic molecule
- Secondary structure prediction algorithms
- RNA covariant analysis

RNA

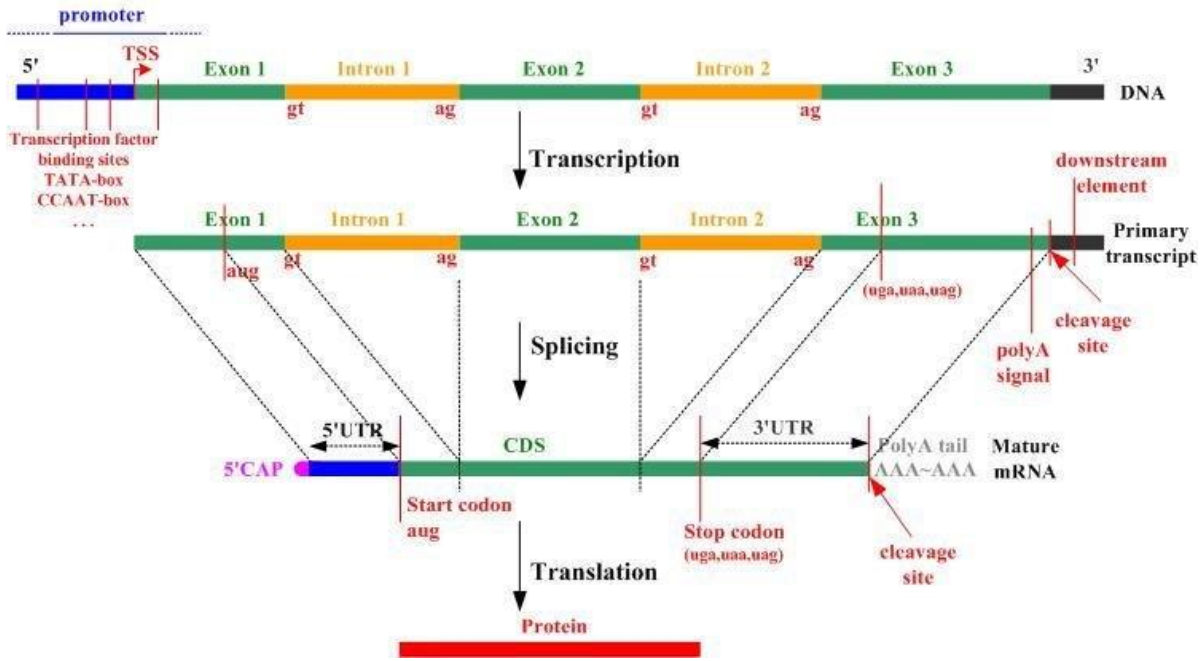
RNA functions as:

- Messenger RNA (mRNA): linear, unstructured sequenced, coding protein amino acid sequences.
- Ribosomal RNA (rRNA)
- Transfer RNA (tRNA)
- Small nuclear RNA (snRNA): component of splicesome
- Small nucleolar (snoRNA): takes part in processing rRNA
- Small interfering RNA (siRNA)
- Micro RNA (miRNA)
- Piwi RNA (piRNA)
- Guide RNA (gRNA)
- Enzymes (ribozymes)
- Viral genomes
- Retrotransposons

Non-coding *vs.* coding RNA

- Non-coding RNA (ncRNA):
  - RNA that is not translated into protein.
  - Including: tRNA, rRNA, snRNA, snoRNA, miRNA, gRNA, piRNA,…

- mRNA – coding RNA:
  - 5’ methylated cap
  - 5’ – UTR (un-translated regions)
  - CDS (coding sequence)
  - 3’ – UTR
  - Poly-A tail
  - mRNA untranslated regions (UTRs) are not considered ncRNA



RNA 101

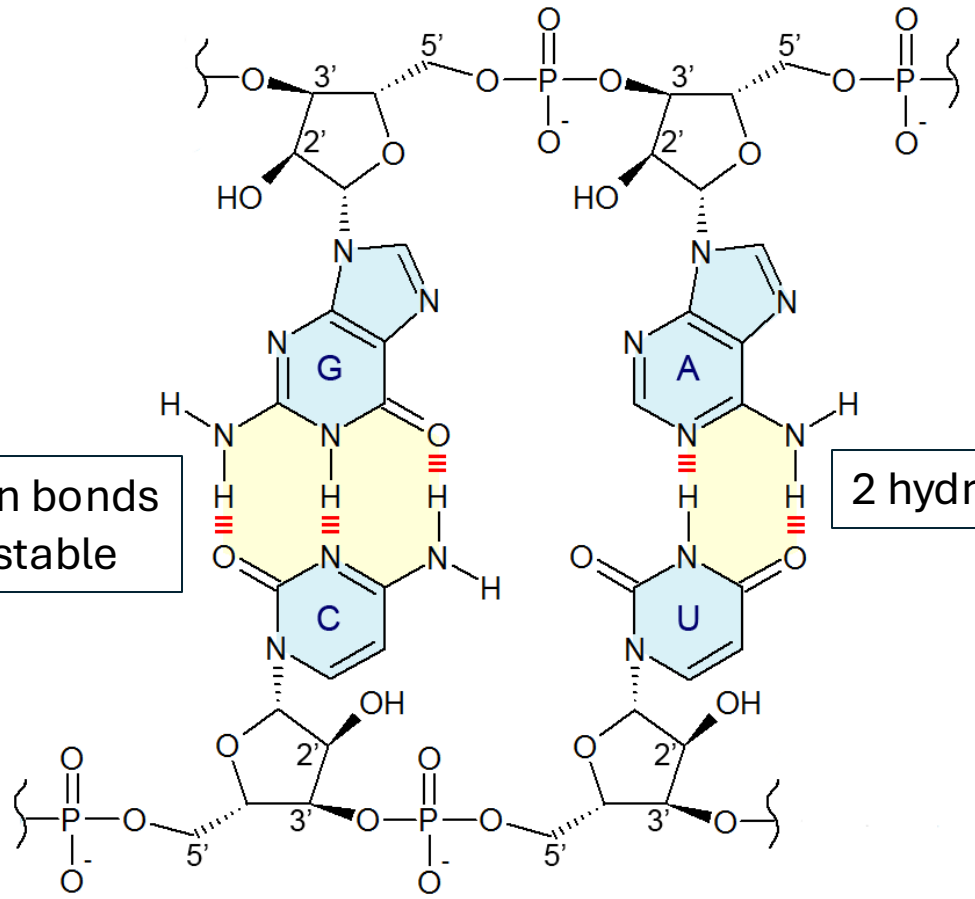
- RNA bases: adenine (A), guanine (G), uracil (U), and cytosine (C).

- Watson-Crick Pair
  - A-U (~ 2 kcal/mol)
  - G-C (~ 3 kcal/mol)

- Non-Canonical pairs:
  - Wobble pair: G-U (~ 1 kcal/mol)

- Bases can only pair with one other base

3 hydrogen bonds = more stable



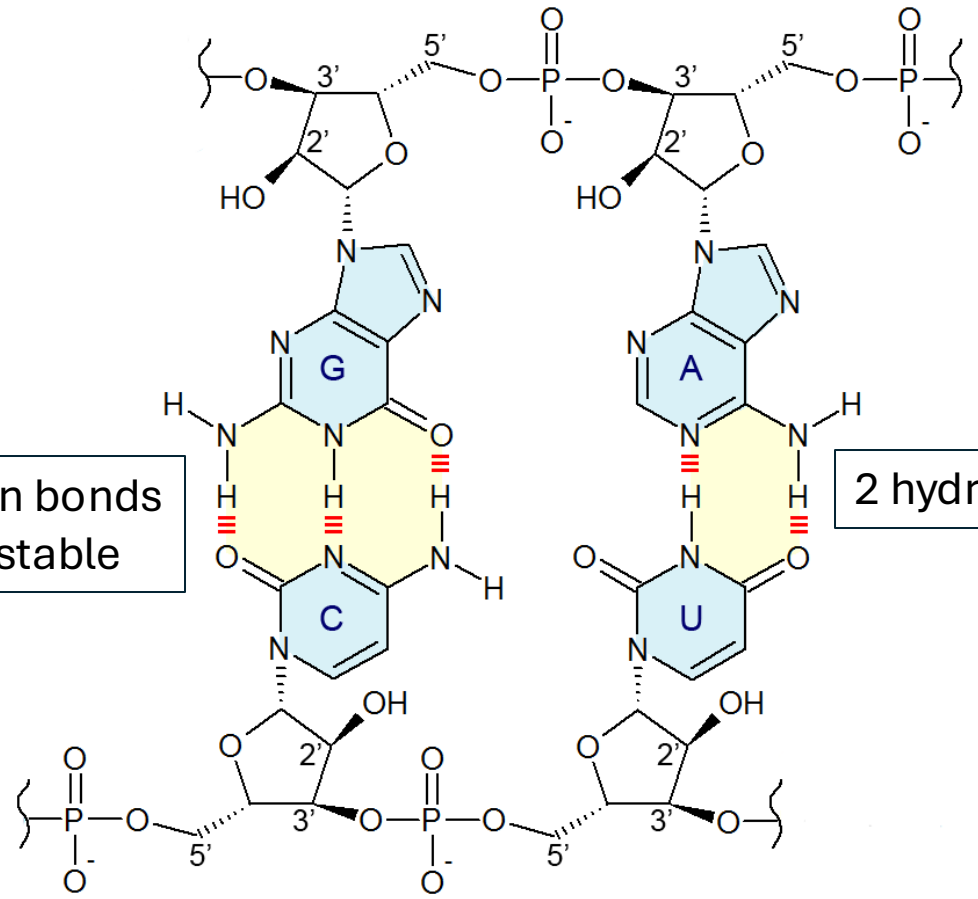
2 hydrogen bonds

Source: wikipedia

RNA 101

- Uracil instead of Thymine
- Sugar is ribose instead of Deoxyribose
- DNA is primarily in duplex form
- RNA is Single stranded: can have a diverse form of secondary structures, other than duplex

3 hydrogen bonds = more stable

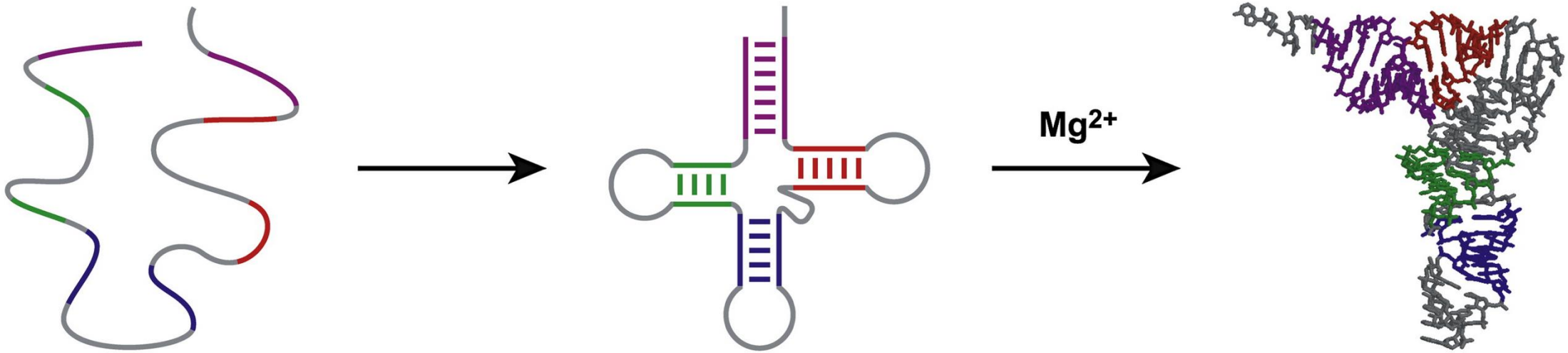


2 hydrogen bonds

Source: wikipedia

RNA 101

Yeast tRNA - Phe

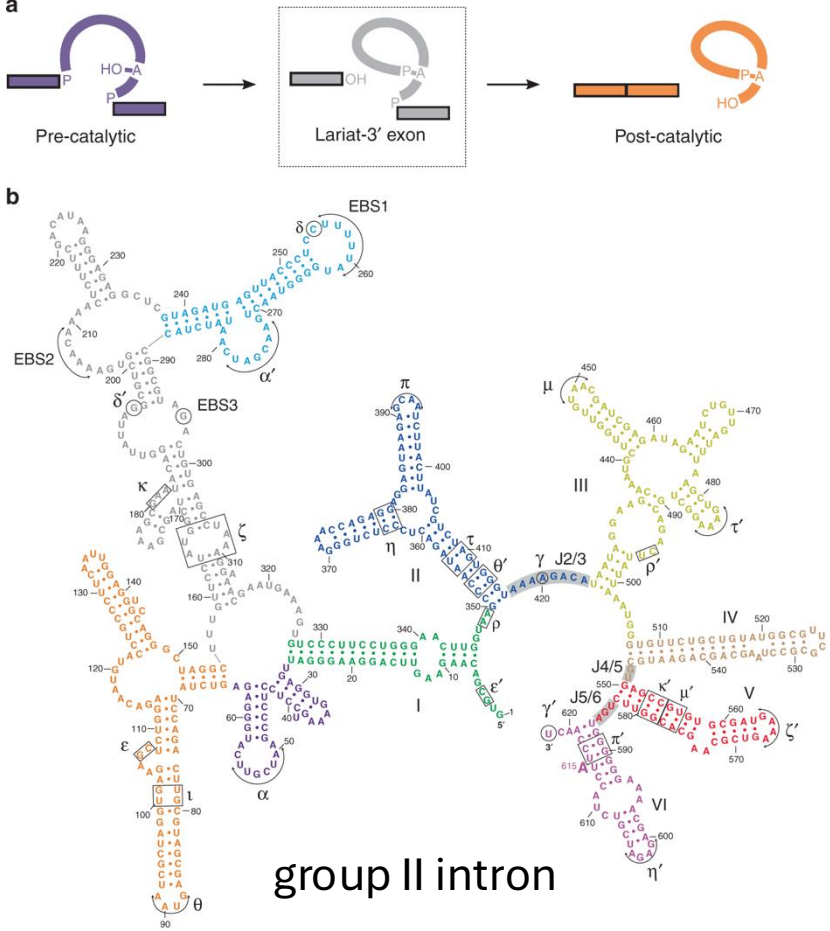


primary (1°) sequence

secondary (2°) structure

tertiary (3°) structure

RNA 101

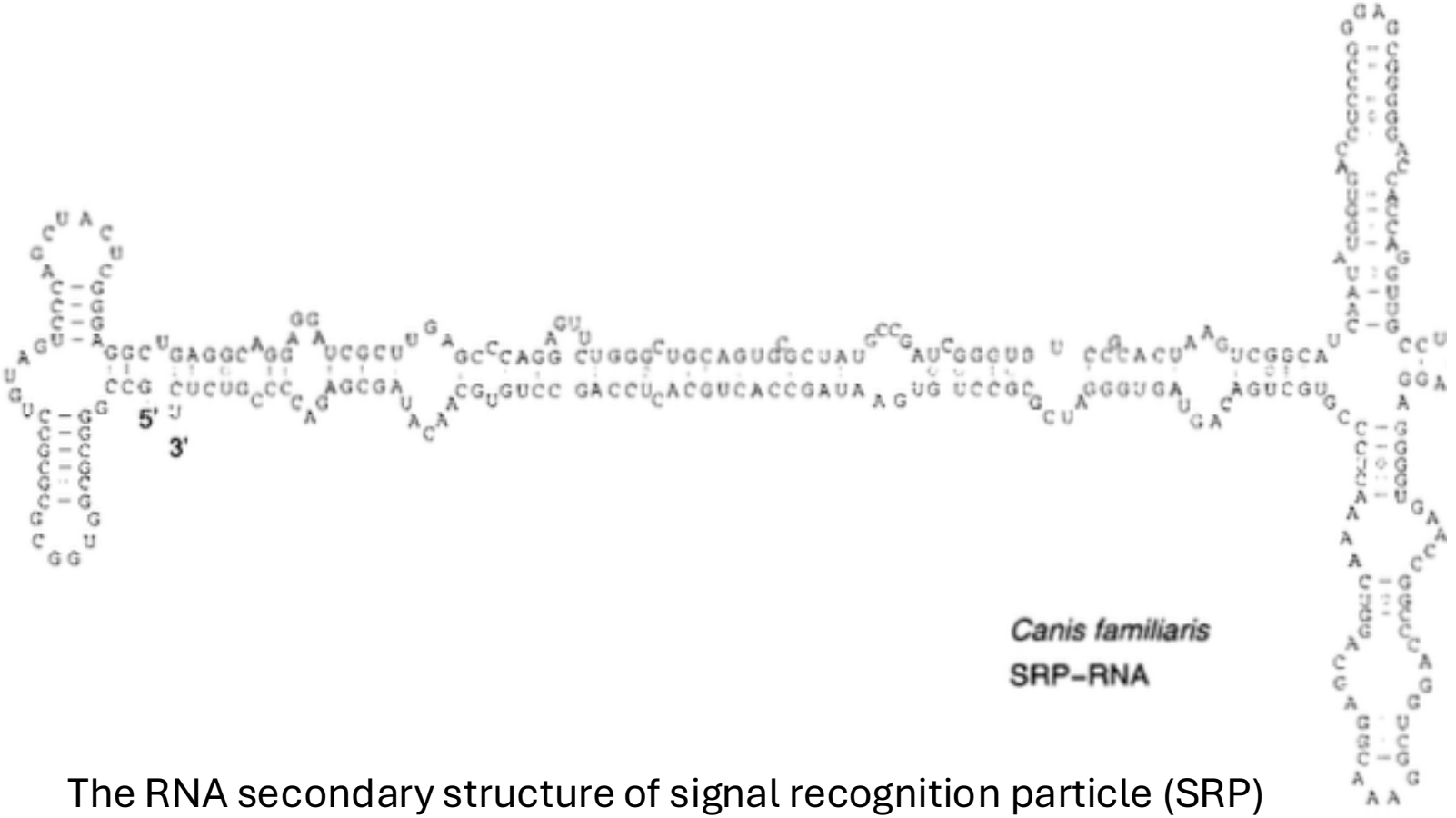


Source: https://doi.org/10.1038/

<u>Some structural rules:</u>

- Base pairing is stabilizing.
- Unpaired sections (loops) destabilize.
- RNA 3D conformation with interactions is made up of a complex, folded structure formed by the RNA molecule itself, stabilized by:
  - Base pairing
  - Base stacking (interactions within three-dimensional architectures)
  - Interaction with other molecules

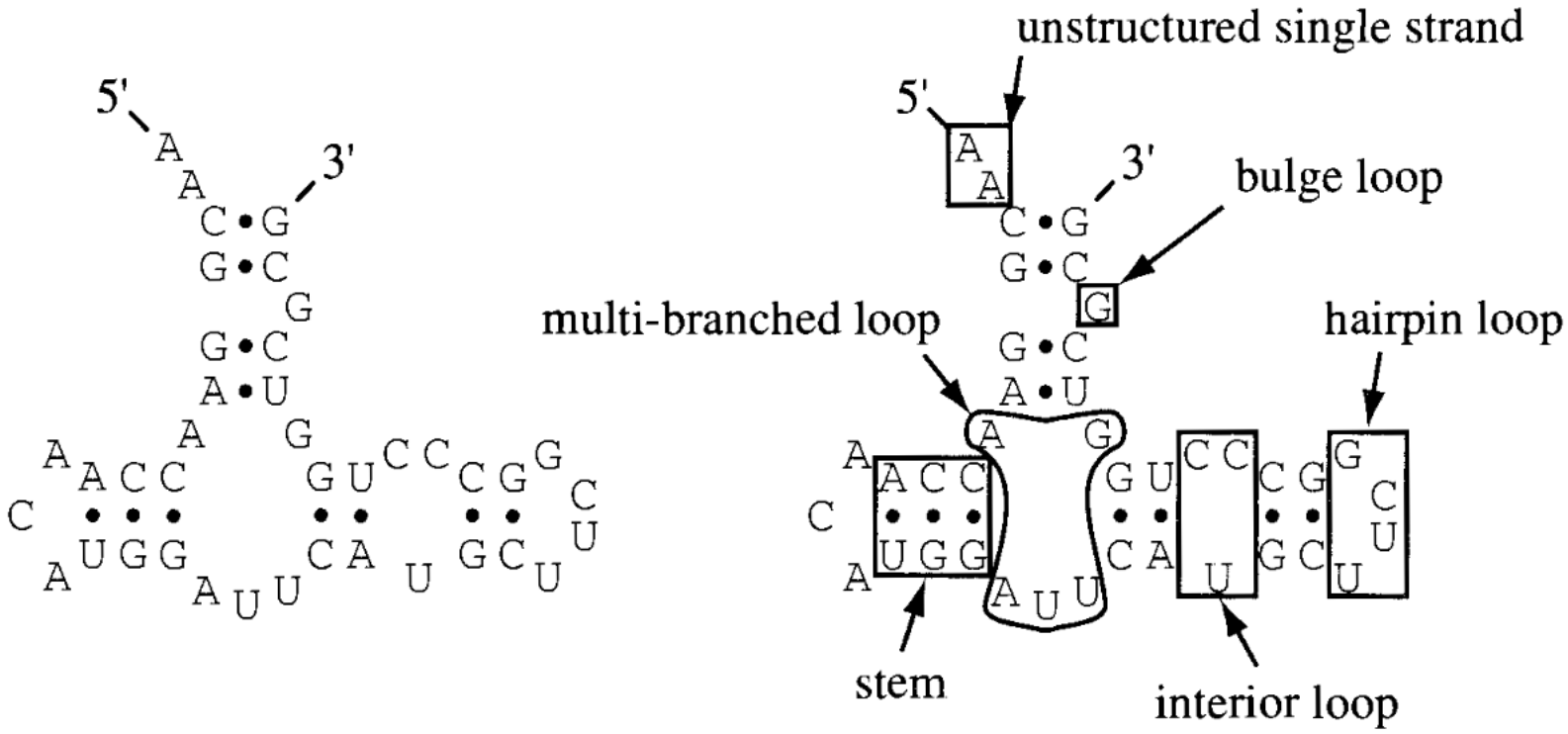
RNA secondary structure elements



The RNA secondary structure of signal recognition particle (SRP) RNA from the dog, *Canis familiaris*.

Source: *Biological sequence analysis*, Durbin *et al.*

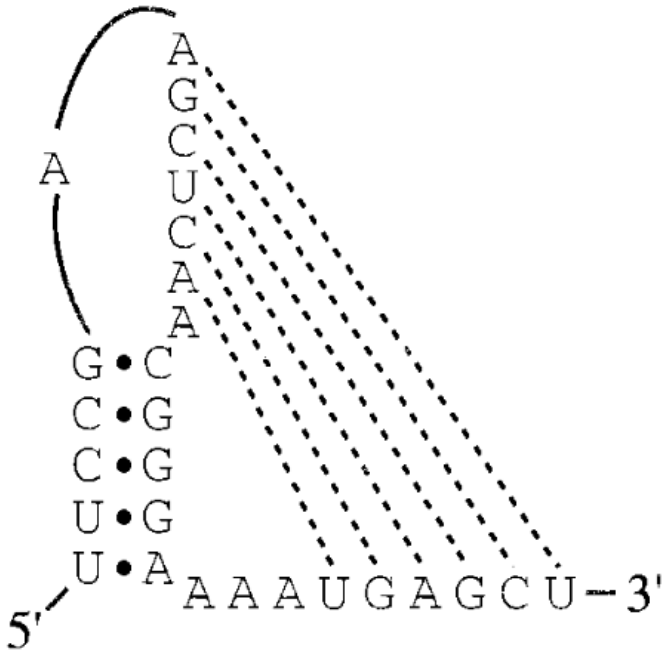
RNA secondary structure elements

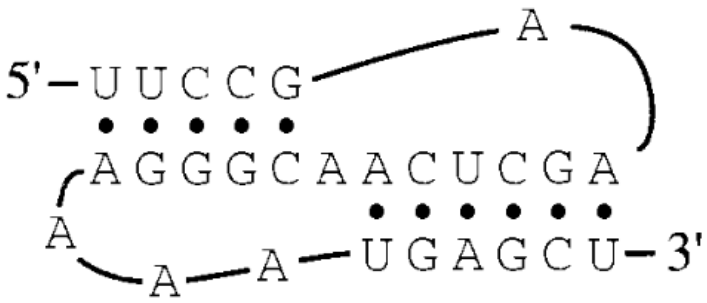


Fundamental elements of RNA secondary structure (hypothetical example).

Source: *Biological sequence analysis*, Durbin *et al.*

RNA secondary structure elements



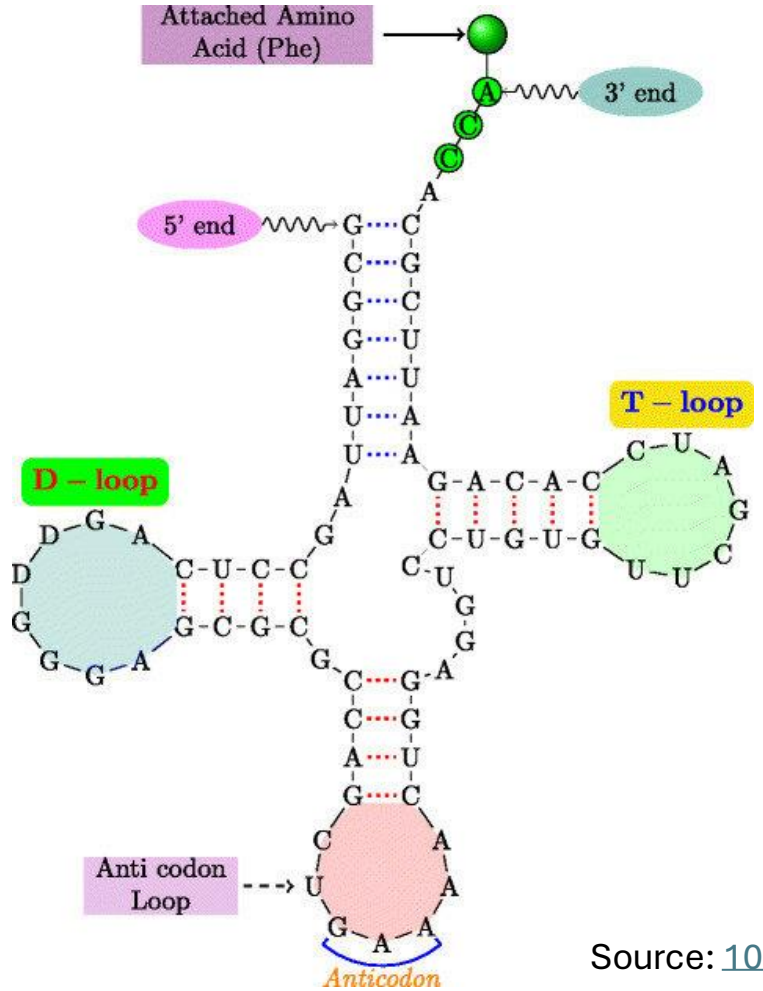


Source: *Biological sequence analysis*, Durbin *et al.*

When non-nested base pairs occur, they are called **pseudoknots**: base pairs between a loop and positions outside the enclosing stem.

- None of the dynamic programming algorithms for RNA folding (Nussinov and Zuker) can deal with pseudoknots.
- Pseudoknots occur in many important RNAs.
- Although challenging to deal with them, the total number of pseudoknotted base pairs is relatively small.

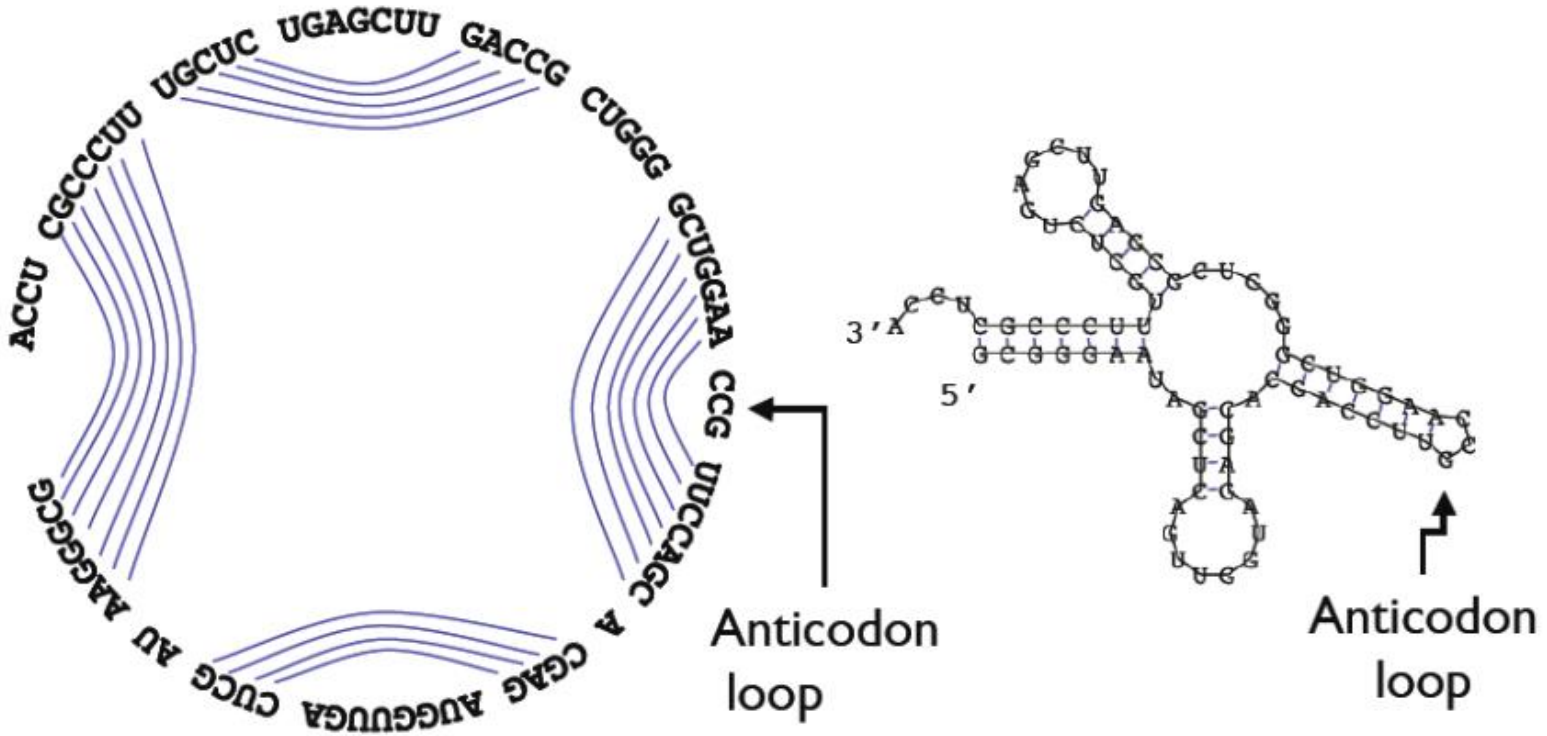
tRNA representations



Secondary structure of Yeast tRNA Phenylalanine

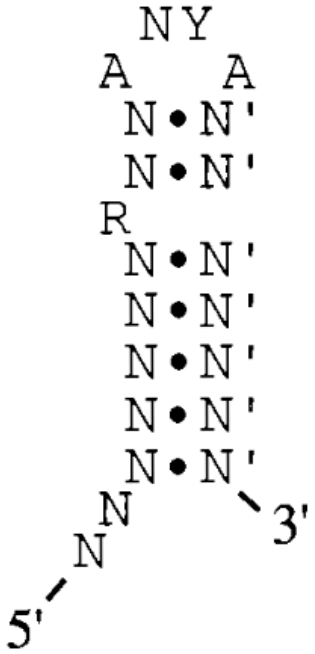
Source: 10.1186/s13062-015-0090-5

tRNA representations



RNA sequence

- RNA sequence evolution is constrained by structure.
- It is common to find examples of homologous RNAs that have a similar secondary structure without sharing significant sequence similarity.
- In this case, the standard sequence alignment method is not very useful in searching homologs.



Consensus binding site for R17 phage coat protein.

N = {A, C, G, U}

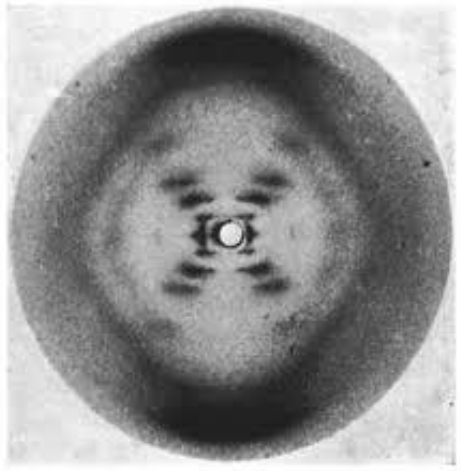
Y = {C, U}

R = {A, G}

N' = complementary base pairing to N

Why predict structures?

- In RNA, function depends on the sequence and secondary structure: protein binding, base pairing to another RNA, modifying a nucleic acid bond.
- Knowing the shape of any biomolecule is of great value in understanding its mechanism.
- Current physical methods (X-ray, NMR) are too expensive and time-consuming.
- We need to predict RNA shape from the sequence of bases.



Approaches to structure prediction

- Maximum pairing
  - Works on single sequences
  - It is simple
  - But it is too inaccurate

- Minimum energy
  - Works on single sequences
  - Ignores pseudoknots
  - Only find the *optimal* fold

- Comparative sequence analysis
  - Handles all pairings (pseudoknots included)
  - Required several aligned, appropriately divergence sequences

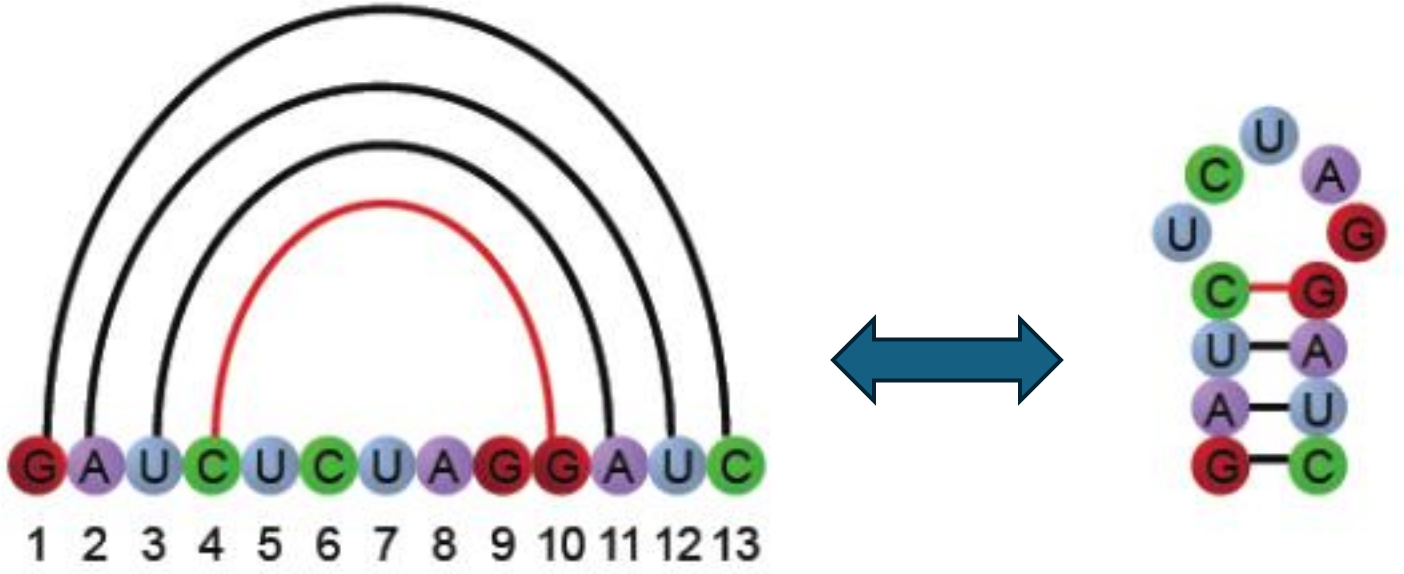
- Physical experiments (X-ray, NMR)

Structure representation

- The secondary structure is described as a graph
- Base pairs are described via pairs of indices (i, j), indicating links between base vertices:

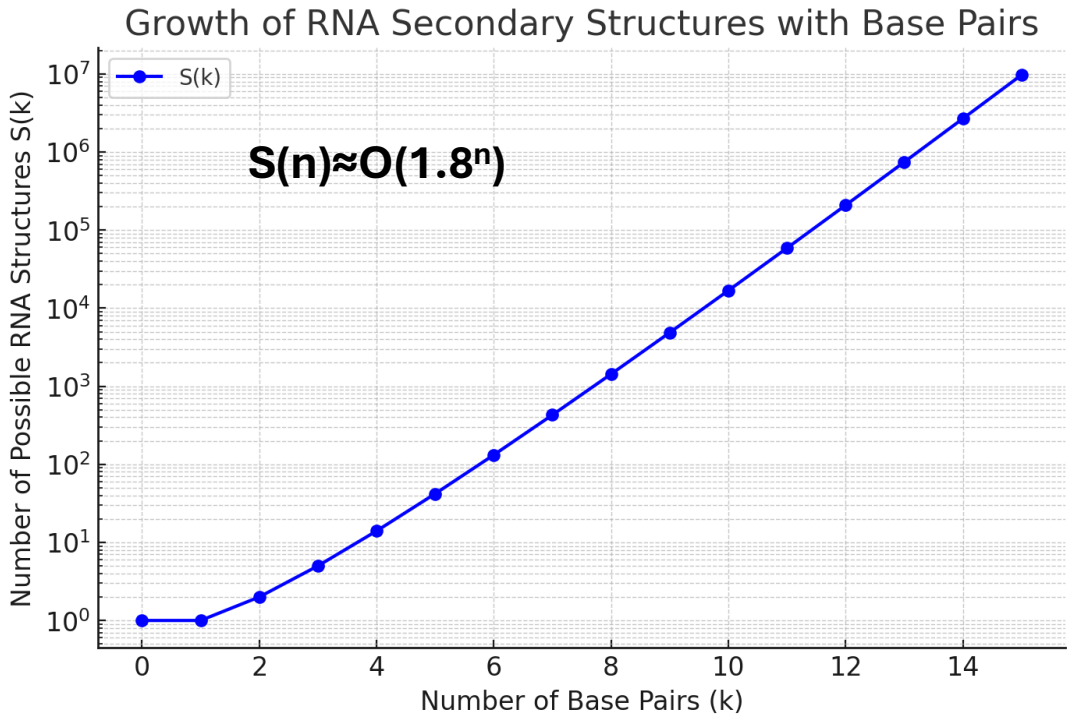
Stem (S)

$S = \{(1, 13), (2, 12), (3, 11), (4, 10)\}$



Possible secondary structures

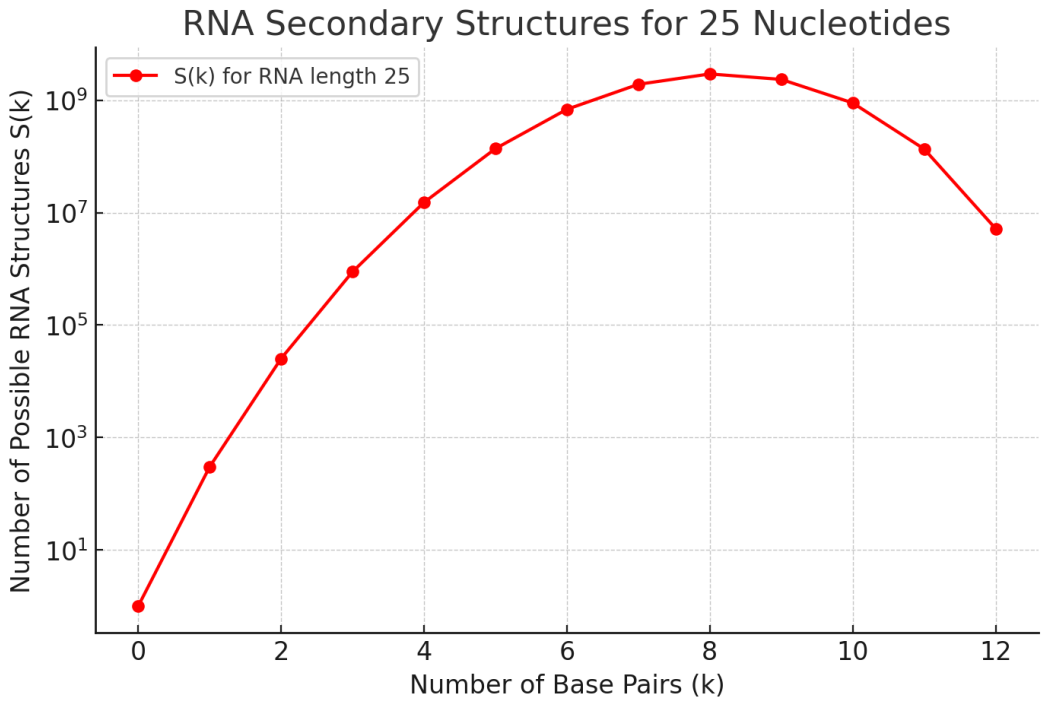
- k = number of base pairs
- S = number of possible secondary structures



Plot showing the number of possible RNA secondary structures S(k) as a function of the number of base pairs k. The *y-axis* is in logarithmic scale to better visualize the rapid growth.

Possible secondary structures

- n = 25 bases long
- k = number of base pairs
- S = number of possible secondary structures



The y-axis is in logarithmic

Peak at 8 base pairs = $2.9*10^9$ possible base paired structures

An RNA of only 90 bases long -> 10^{39} !!!

An RNA of only 200 bases long -> 10^{94} !!!

Not feasible using Brute Force!

Nussinov algorithm

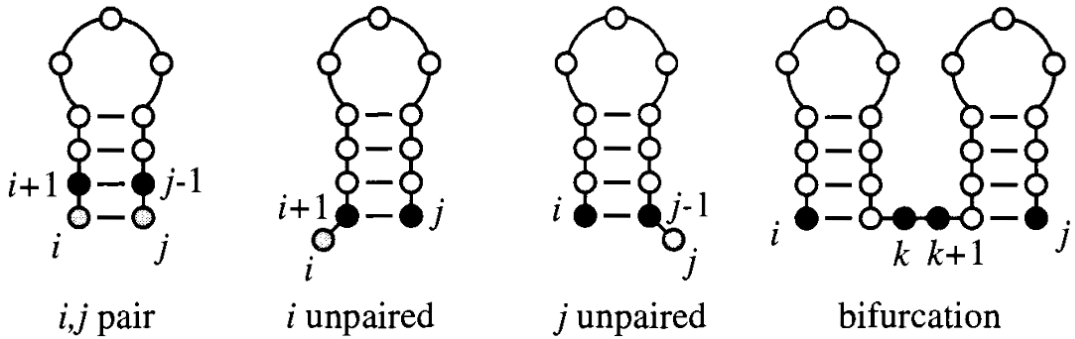
The Nussinov folding algorithm finds the structure with the most base pairs. It is an efficient dynamic programming algorithm (Nussinov et al. 1978) with simplistic criteria that gives accurate structure predictions.

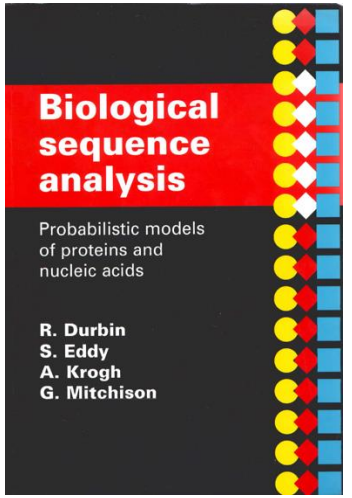
- Assumes that all base pairs have equal thermodynamic energy.
- Compares a sequence against itself in a **dynamic programming** matrix.
- It finds local regions of continuous base pairings.
- The Nussinov calculation is **recursive**. It calculates the best structure for small subsequences and works its way outwards to larger and larger subsequences.
- Nussinov dynamic programming algorithm requires $O(n^3)$ steps and $O(n^2)$ storage.

Nussinov algorithm

The main idea of the recursive calculation is that there are only four possible ways of getting the best structure for *i* and *j* from the best structures of the smaller subsequences.

1. Add unpaired position *i* onto best structure for subsequence $i + 1, j$
2. Add unpaired position *j* onto best structure for subsequence $i, j - 1$
3. Add *i*, *j* pair onto best structure found for subsequence $i + 1, j - 1$
4. Combine two optimal substructures, *i*, *k* and $k + 1, j$

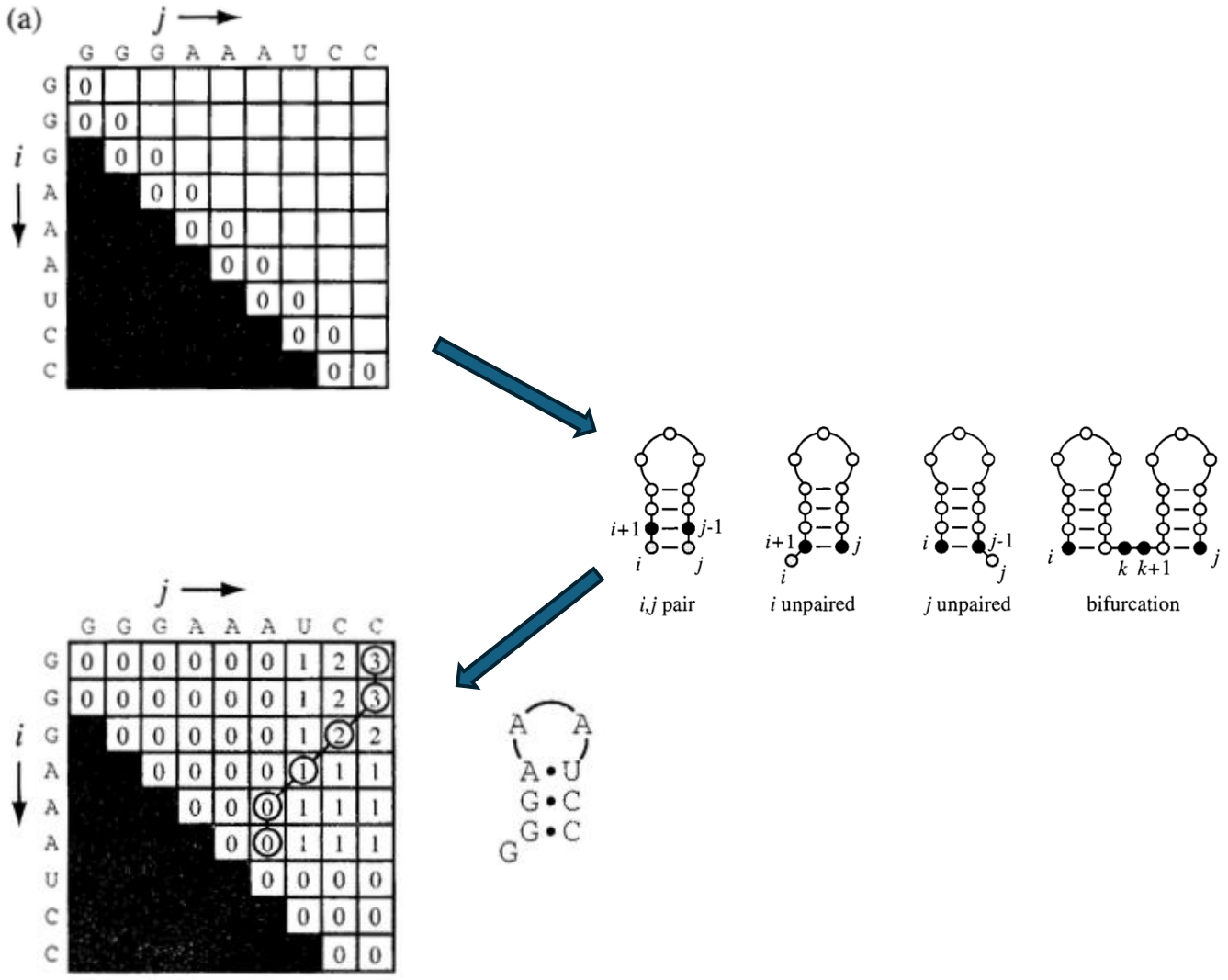




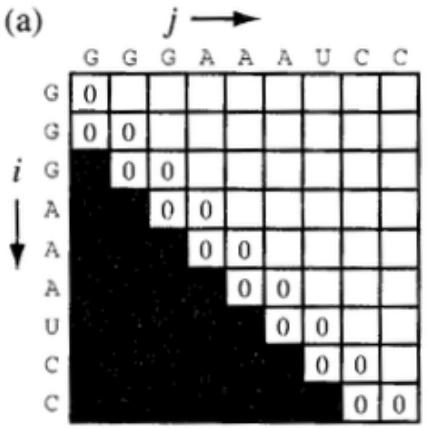
Nussinov algorithm

Stages:

- Fill stage
  - Initialization
  - Extension (Recursion)
- Traceback



Nussinov initialization



Initialization:

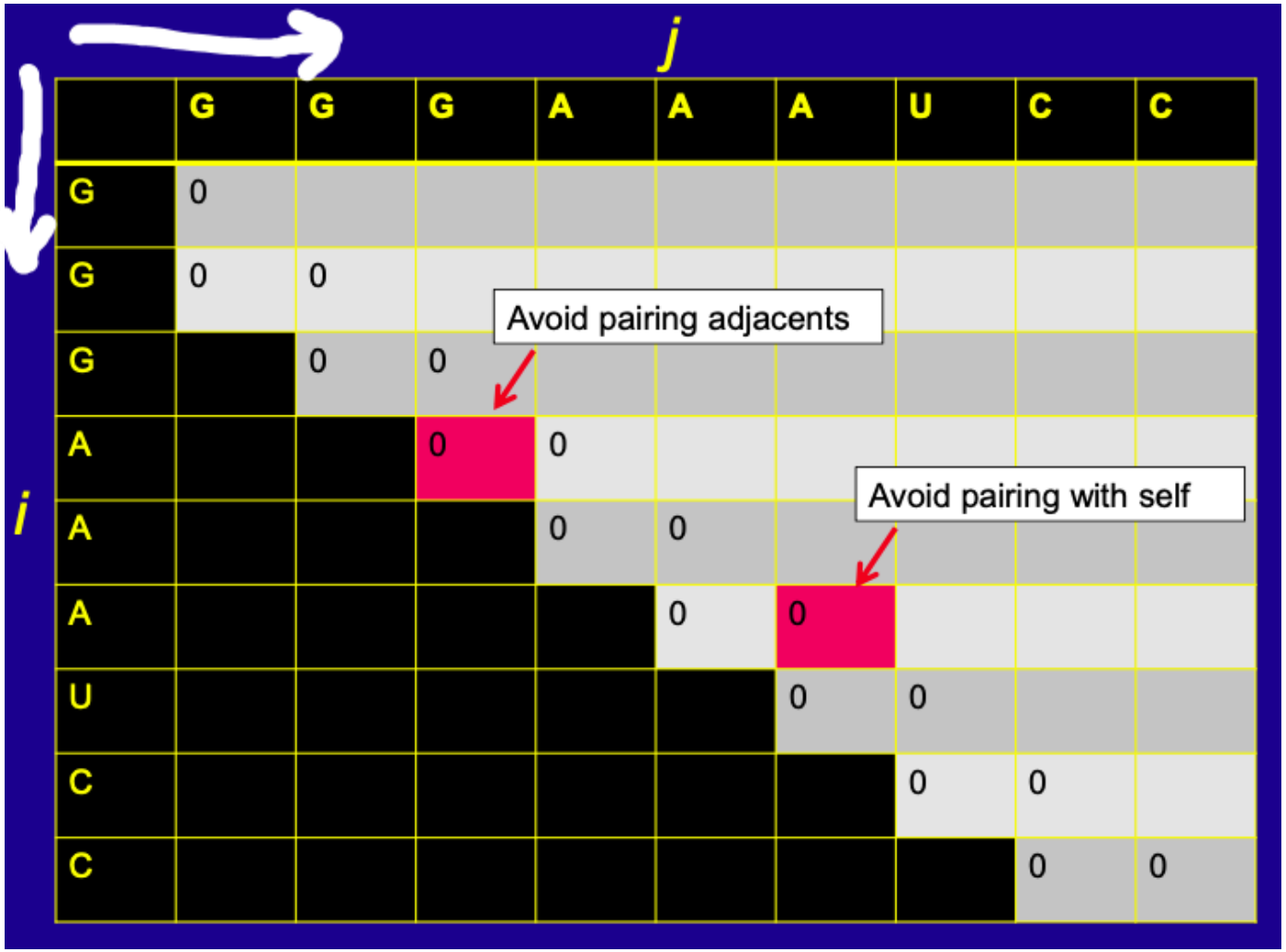
- In the matrix, each half is redundant with the other.
- The score for matches along the main diagonal and diagonal just below are set to zero.
  - No pairing with self.
  - No pairing (preferably) with adjacent nucleotides.

Initialisation:

$$\gamma(i, i-1) = 0 \qquad \text{for } i = 2 \text{ to } L;$$
$$\gamma(i, i) = 0 \qquad \text{for } i = 1 \text{ to } L.$$

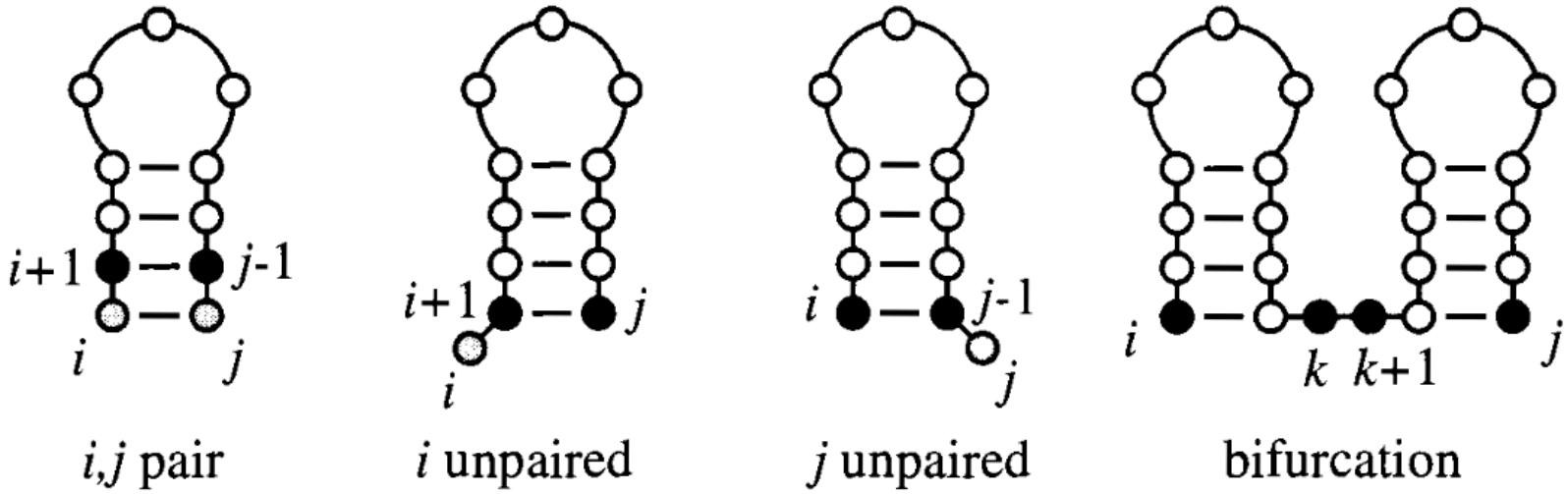
Source: *Biological sequence analysis*, Durbin *et al.*

Nussinov initialization



Nussinov extension

four possible ways of getting the extension



The Nussinov algorithm looks at four ways in which the best RNA structure for a subsequence can be made by adding onto already calculated optimal structures for smaller subsequences.
* Pseudoknots are not considered.

Source: *Biological sequence analysis*, Durbin *et al*.

Nussinov extension

Extension:

- In the matrix, for each cell, it is tested the four possibilities.
- We make use of the recursion:

Recursion: starting with all subsequences of length 2, to length L:

$$\gamma(i,j) = \max \begin{cases} \gamma(i+1,j), & \text{unpaired} \\ \gamma(i,j-1), & \text{unpaired} \\ \gamma(i+1,j-1)+\delta(i,j), & \text{paired} \\ \max_{i<k<j}\left[\gamma(i,k)+\gamma(k+1,j)\right]. & \text{bifurcation} \end{cases}$$

Source: *Biological sequence analysis*, Durbin *et al.*

Nussinov extension

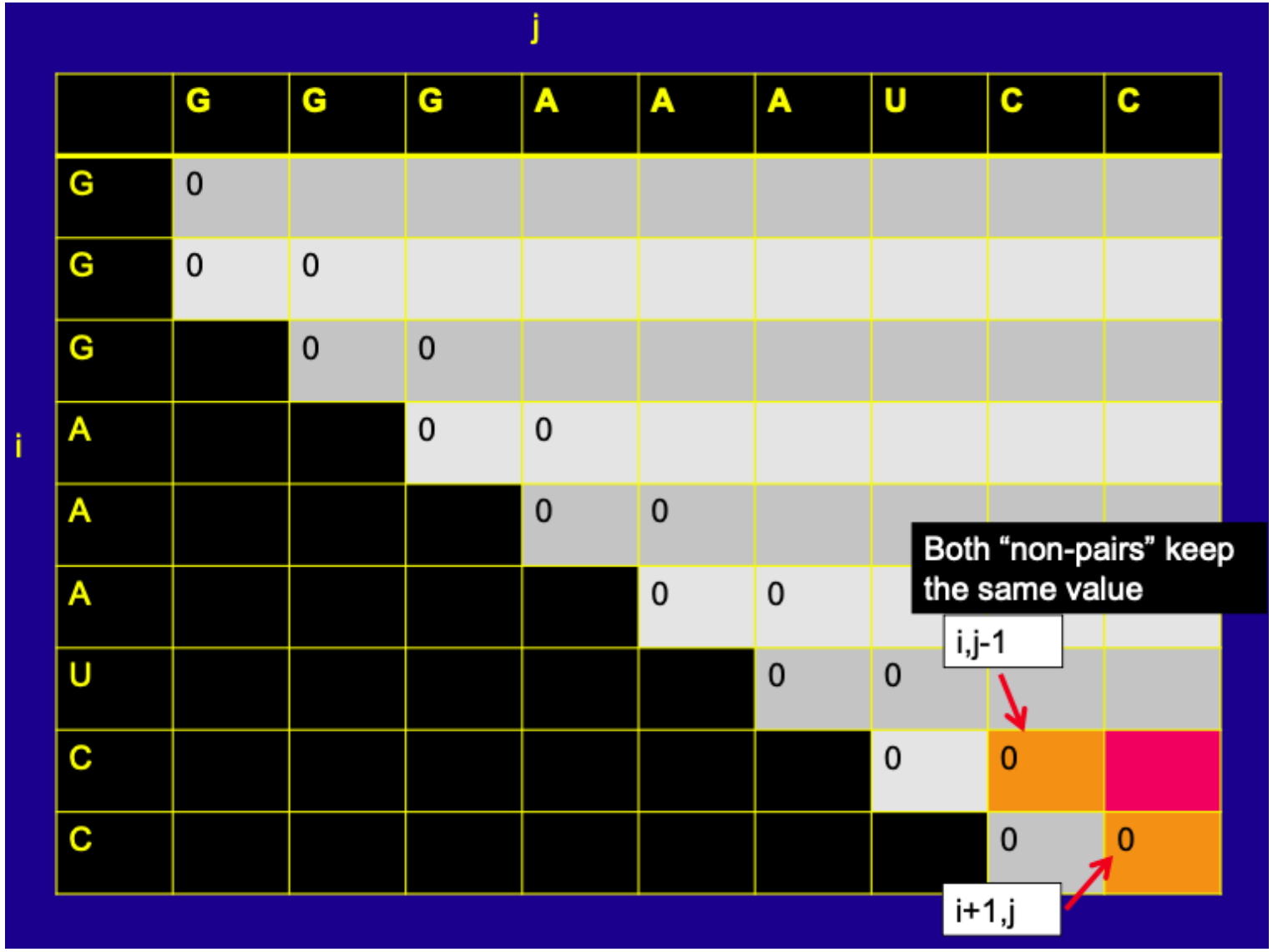
$$\gamma(i,j) = \max \begin{cases} \gamma(i+1,j), \\ \gamma(i,j-1), \\ \gamma(i+1,j-1)+\delta(i,j), \\ \max_{i<k<j}\left[\gamma(i,k)+\gamma(k+1,j)\right]. \end{cases}$$

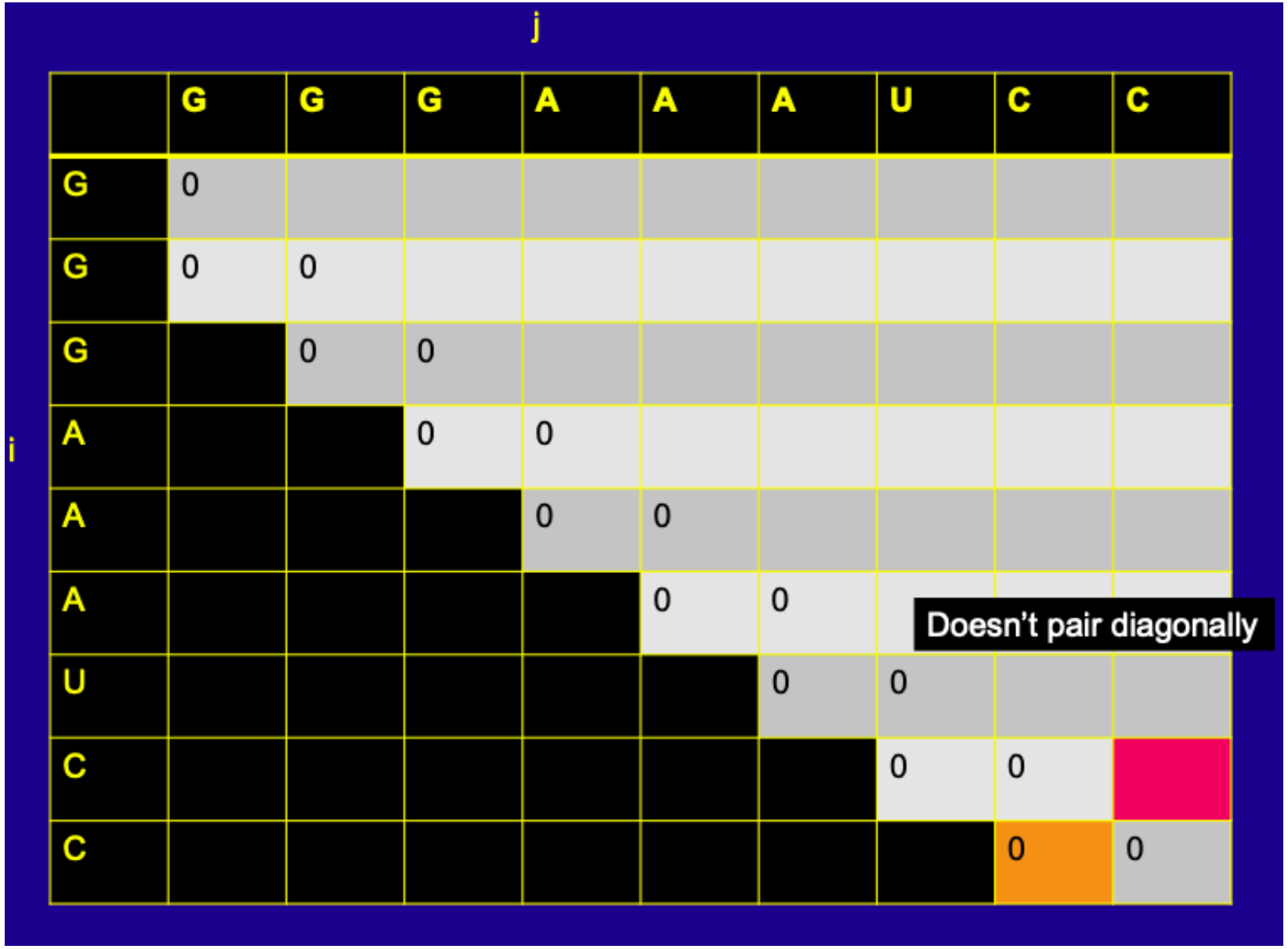
Both of the first two keep the same value from the left or below. No extension in number of base pairs.

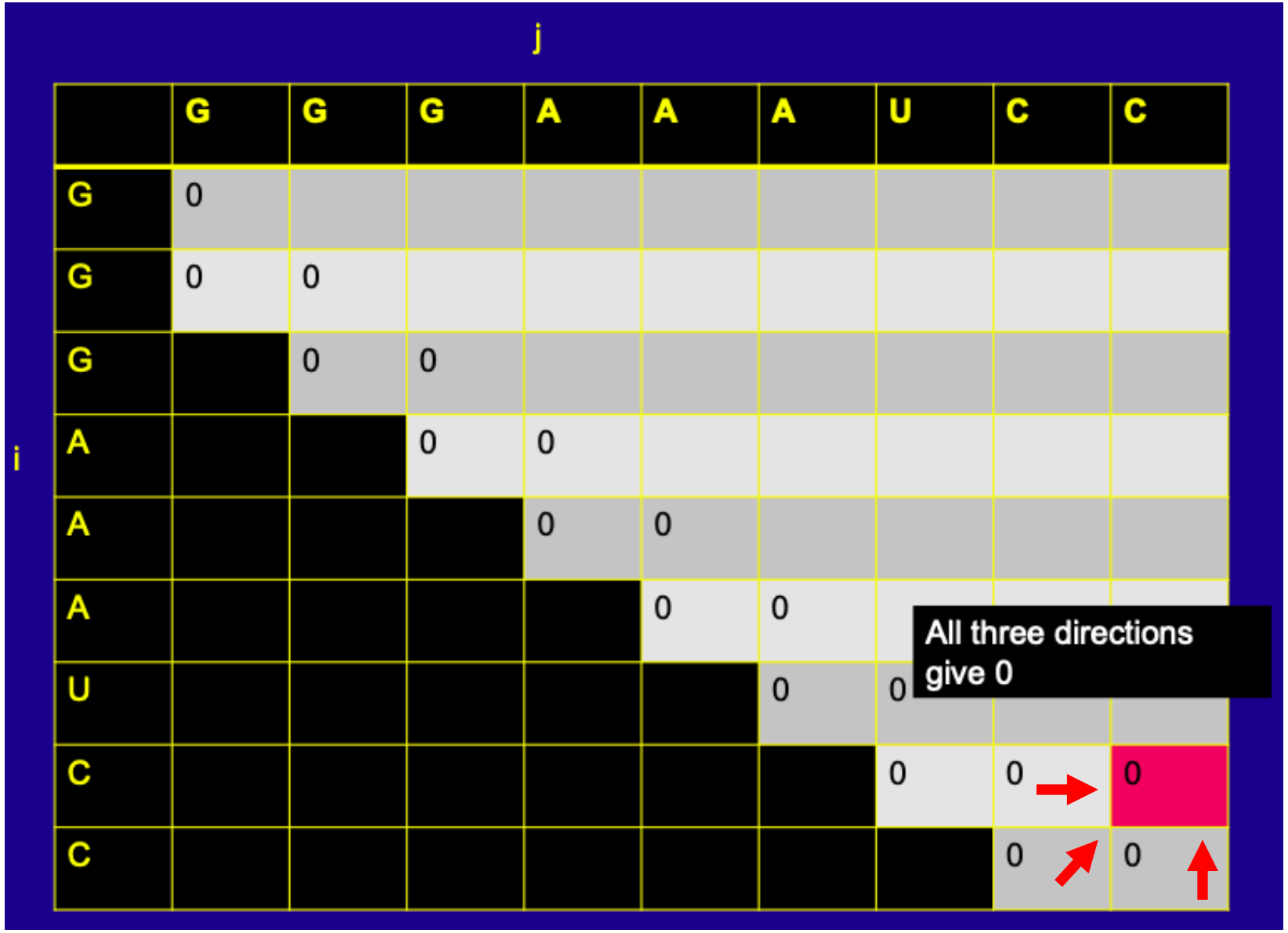
If bases are complimentary: $\delta(i, j) = 1$

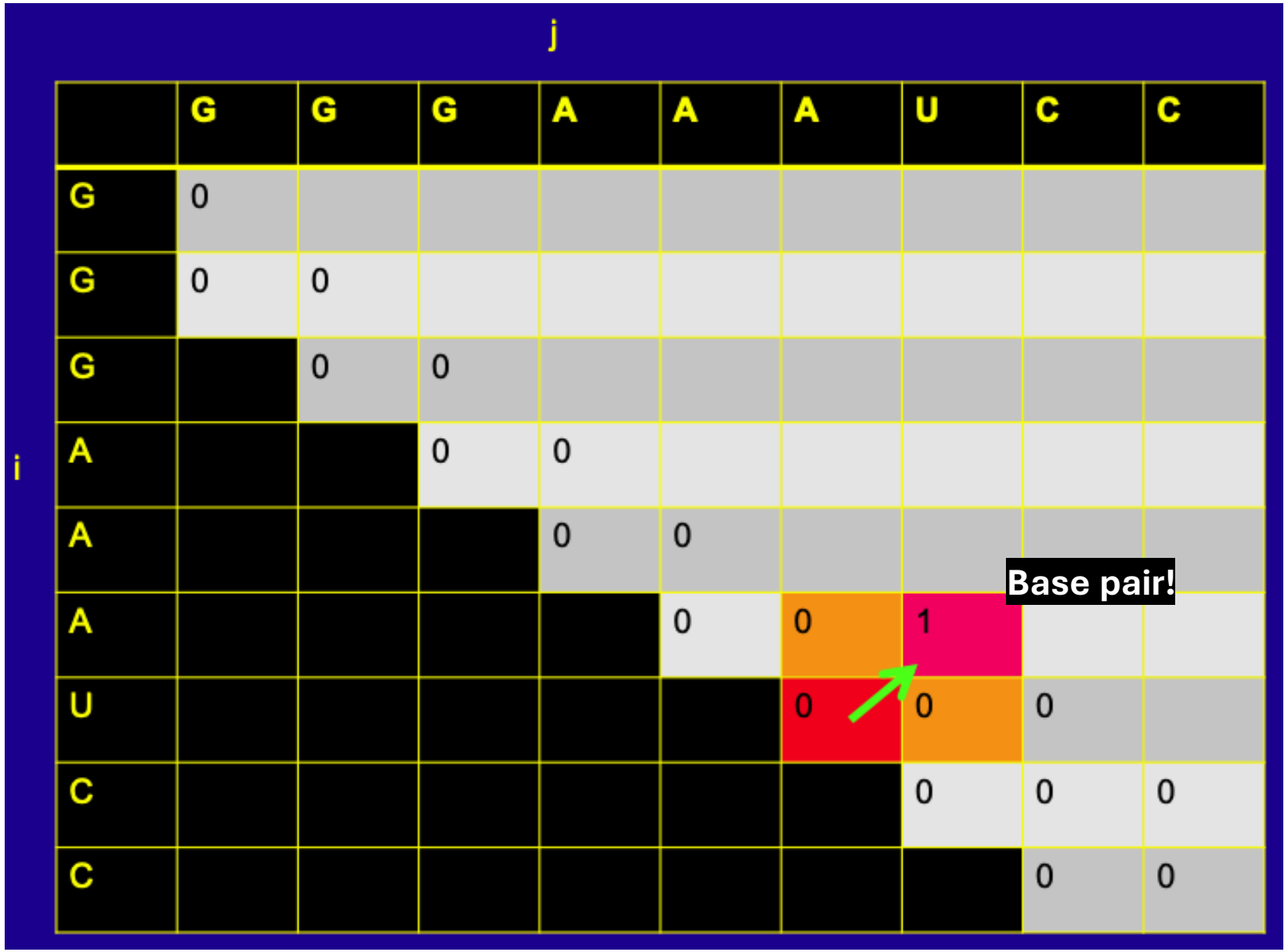
Across all subsequences between *i* and *j*, if there are two that are combined, get a higher score.

Nussinov extension

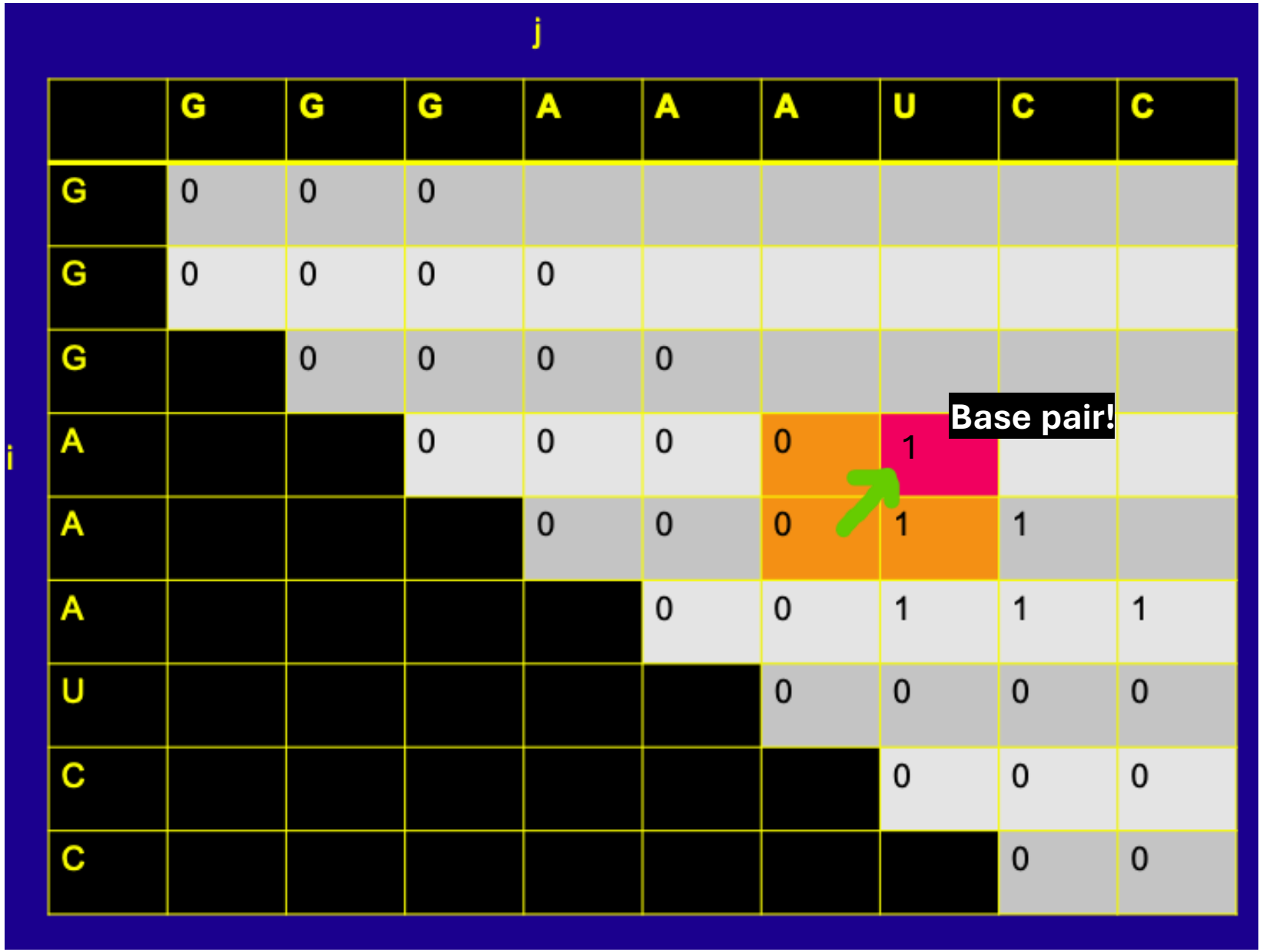



j
i
G G G A A A U C C
G
G
G
A
A
A
U
C
C
0
0 0
0 0
0 0
0 0
0 0
0 0
0 0
0 0
Doesn't pair diagonally

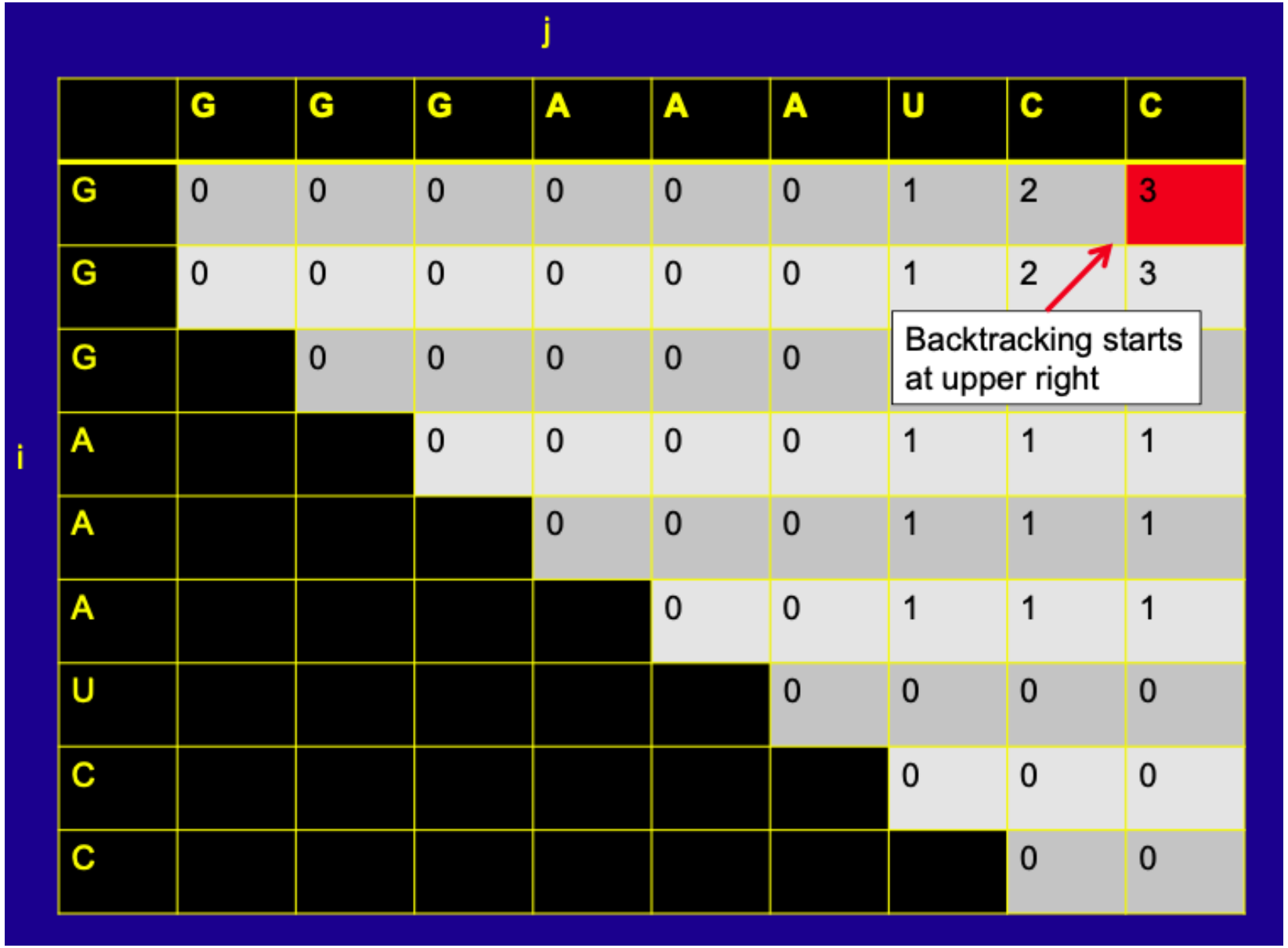

j
i
G G G A A A U C C
G 0
G 0 0
G 0 0
A 0 0
A 0 0
A 0 0
U 0 0
C 0 0 0
C 0 0 0
All three directions give 0



Cell value:
1 (base pair)
+
0 (diagonal)
=
1



Cell value:
1 (base pair)
+
0 (diagonal)
=
1

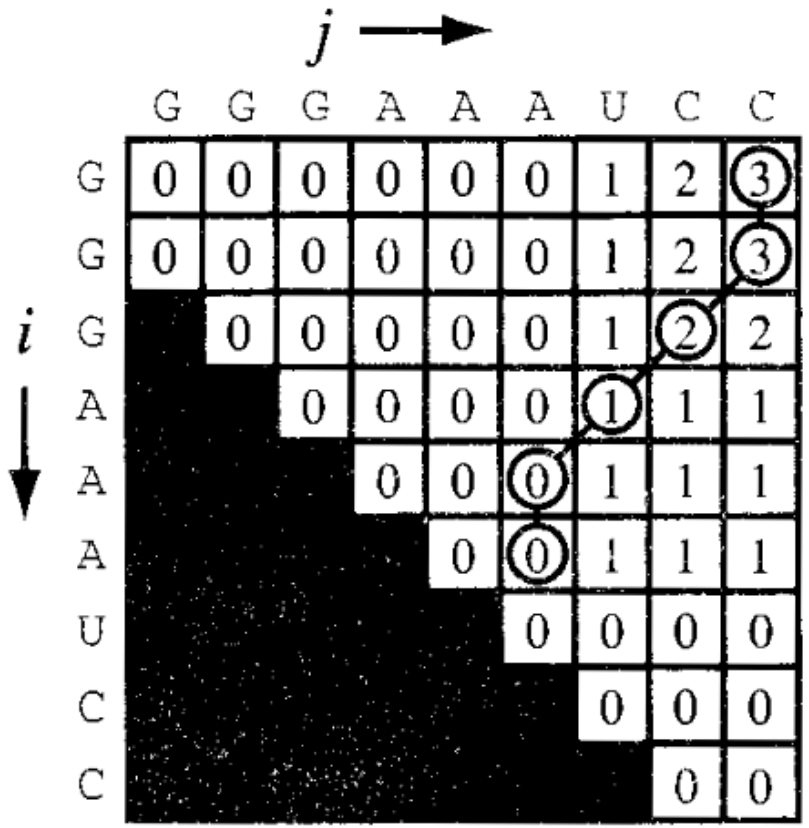


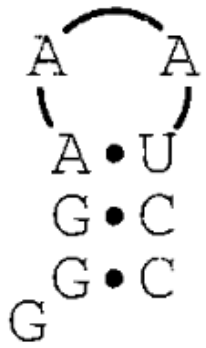
Cell value:
1 (base pair)
+
2 (diagonal)
=
3

Nussinov traceback

- Very similar to pairwise sequence alignment.
- Instead of starting in the lower right corner…
- Go from the **upper right corner** to one of the initialized “rows”

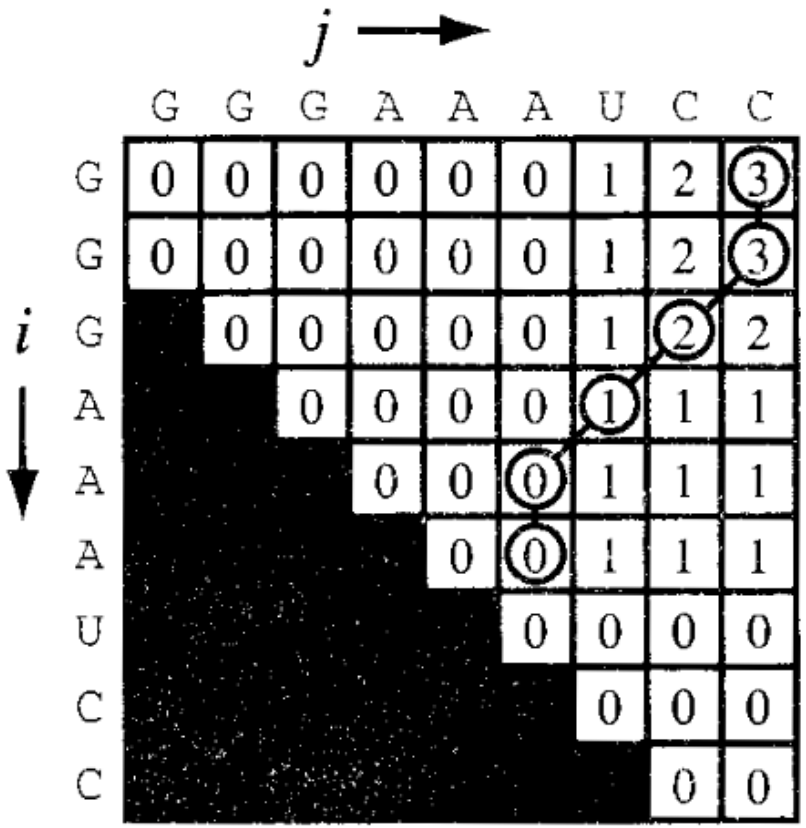
Nussinov traceback



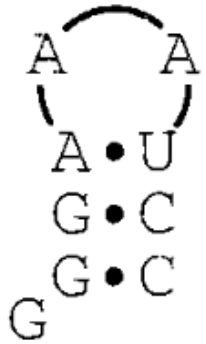


Source: *Biological sequence analysis*, Durbin *et al.*

Nussinov problem

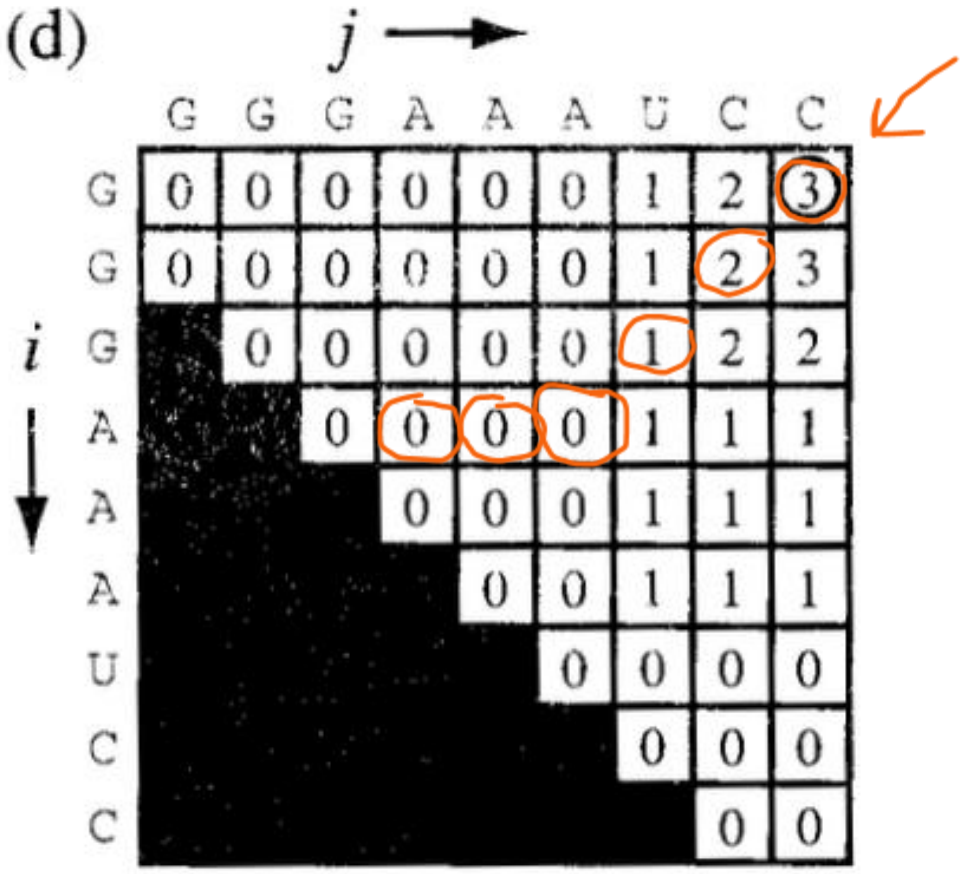


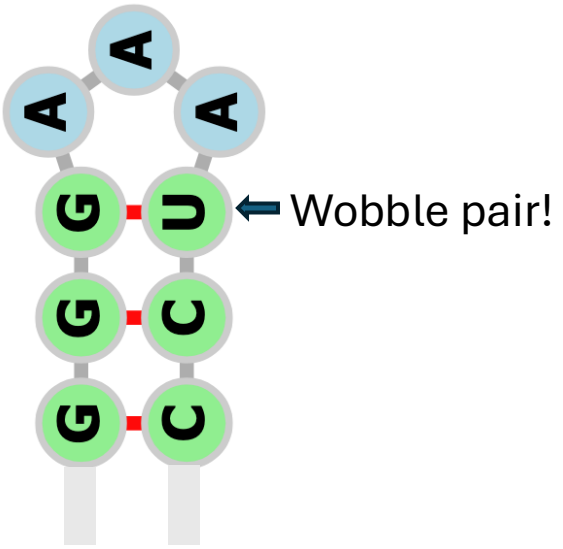
Two-base loop is not structurally stable!



Source: *Biological sequence analysis*, Durbin *et al.*

Nussinov problem





Source: *Biological sequence analysis*, Durbin *et al.*

Nussinov practice

RNA sequence:
GGUCCAC

- Use the Nussinov algorithm to find the maximal number of base pairings
- Via traceback, and visual inspection, find a base paired structure.
- Draw your RNA structure.

Nussinov problem

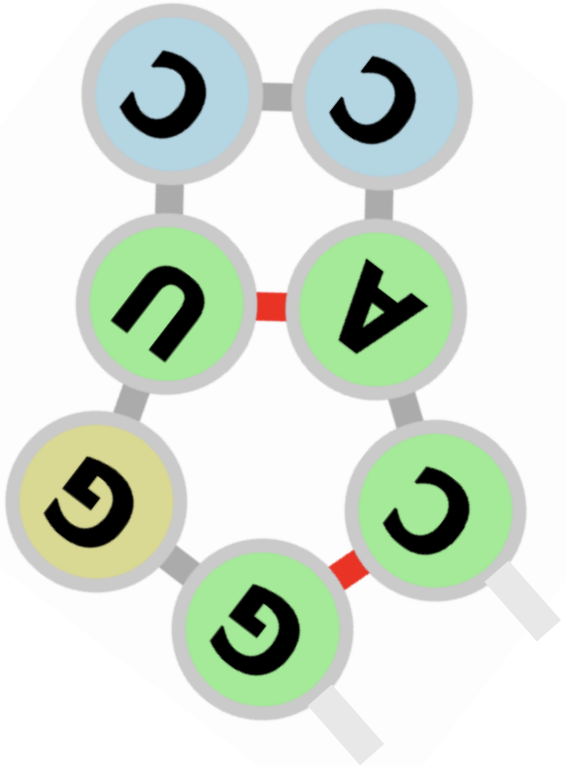
		G	G	U	C	C	A	C
G								
G								
U								
C								
C								
A								
C								

Nussinov problem

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

Nussinov problem

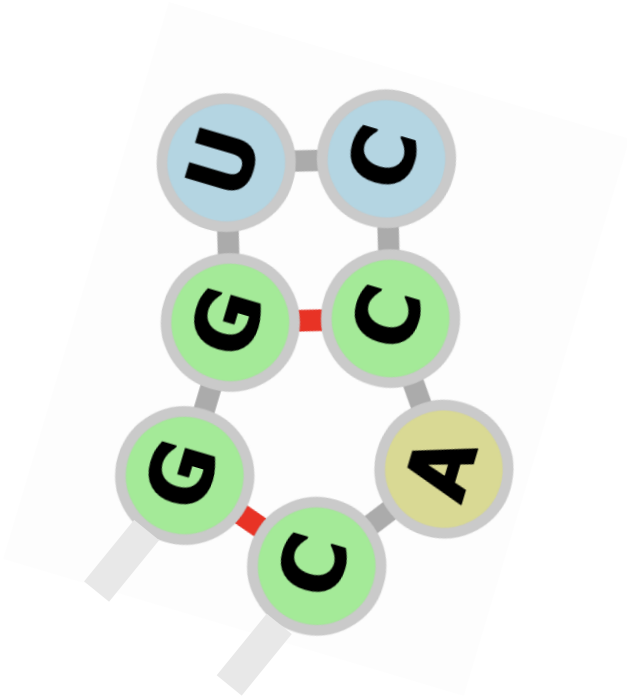
		G	G	U	C	C	A	C
G	0	0	0	1	2	2	2	2
G		0	0	1	1	1	1	2
U			0	0	0	0	1	1
C				0	0	0	0	0
C					0	0	0	0
A						0	0	0
C							0	0



Base pairing

Nussinov problem

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



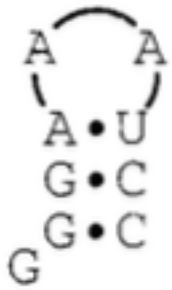
Base pairing

Nussinov modifications

- Nussinov algorithm assumes that all base pairs have **equal thermodynamic energy**.
- But not all base pairs have the exact same free energy: G-C >> A-U>>G-U).
- A structure with **fewer base pairs**, especially those with high pairing probability, **can sometimes have a lower free energy** (more stable - stronger structure) than one with more base pairs.
- Could set $\boldsymbol{\delta}(i, j)$ to reflect free energy of the pairing.

Nussinov modifications

- Base pair maximization will not necessarily lead to the most stable structure.
- It may create a structure with many **interior loops or hairpins ,** which are energetically **unfavorable**.
- Avoiding small loops :
  - Small loops are unstable.
  - At least three unpaired bases in the loop are required.
- Reporting of multiple equivalent scoring structures.
- Cellular conditions affect the stability.
  - Ion concentration
  - temperature



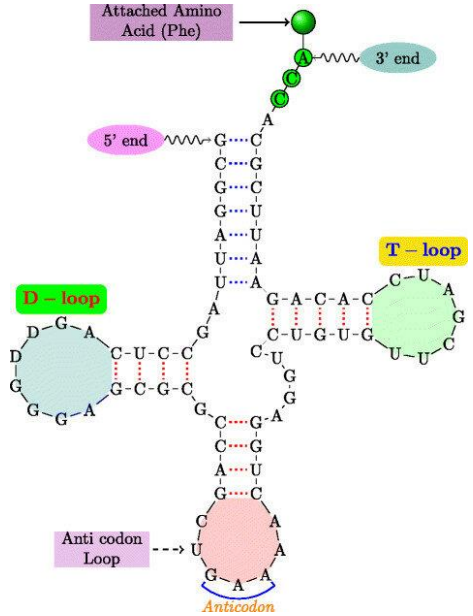
Two-base loop is not structurally stable!

Energy minimization

Zuker algorithm

Looking for Thermodynamic Stability:

- Estimated using experimental techniques.
- Theory: Most Stable = Most likely to exist!
- No pseudoknots due to algorithm limitations.
- Attempts to maximize the score, taking thermodynamics into account.
- Online programs: mFOLD and ViennaRNA
  - http://www.unafold.org/mfold/applications/rna-folding-form.php
  - http://rna.tbi.univie.ac.at/

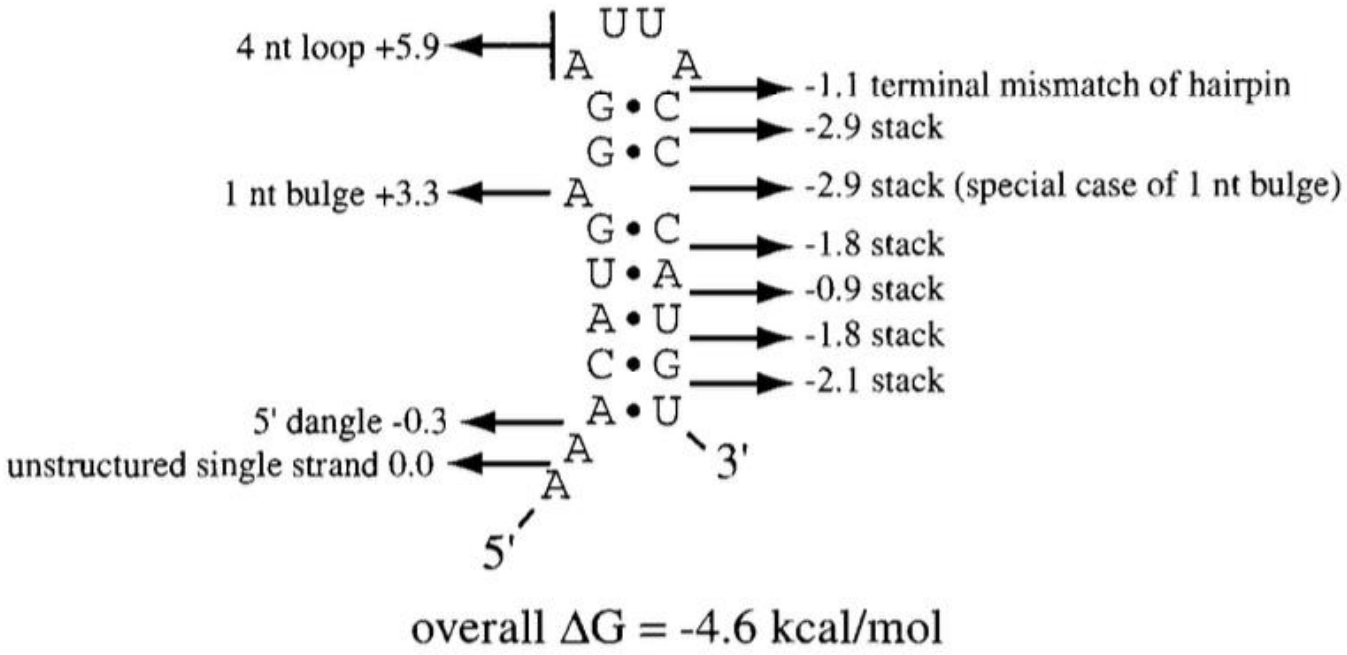


Energy minimization

Zuker algorithm

- Energy is assigned to substructures, not to base pairs.
- Every element has an energy, and the total energy is the sum of all energies.
- The **ΔG** of an RNA secondary structure is approximated as the sum of individual contributions from:
  - Loops
  - Base pairs
  - Bulges
  - Unstructured single strands
  - Stems
- An important difference from Nussinov calculation is that the energy of stems are calculated by adding stacking contributions for the interface between neighboring base pairs instead of individual contributions for each pair.

Zuker algorithm
Minimum free energy calculation



An example of ΔG calculation for an RNA stem loop.

Source: *Biological sequence analysis*, Durbin *et al.*

The hairpin loop energy is the sum of two terms:

-a loop destabilization energy dependent on the loop length.

-a terminal mismatch energy dependent of the closing base pair and the first and last bases of the stem.

Zuker algorithm

- The minimum energy structure (not detailed in BIOL-630) can be calculated recursively by the dynamic programming algorithm.

- There are two matrices to keep track of the stacking interactions.

    - **W**(*i*, *j*) is the energy of the best structure on i, *j*.
    - **V**(*i*, *j*) is the energy of the best structure on *i*, *j* given that *i*, *j* are paired.

- The algorithm can **keep track of stacking interactions** by adding new base pairs only into the V matrix.

- The original Zuker algorithm finds **only the optimal structure**.

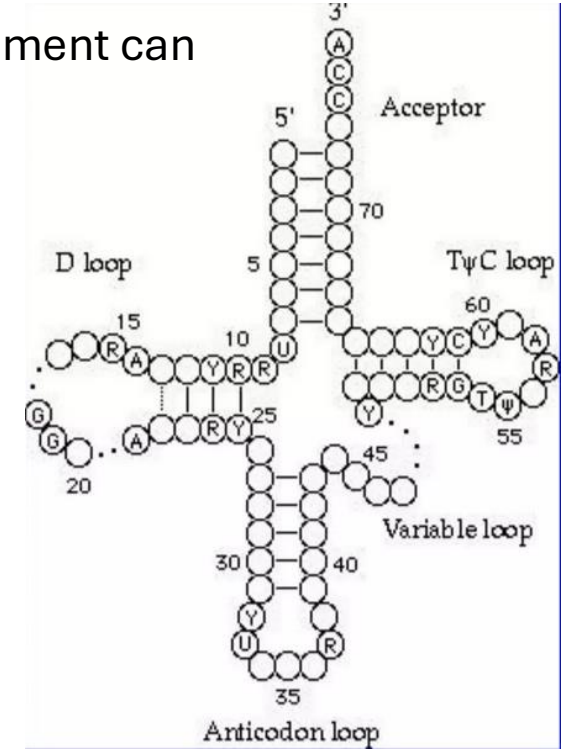
- The optimal solution may not represent the biologically correct solution.

- It does not work with pseudoknots!

Some improvements and implementations of the Zuker algorithm can predict pseudoknots

Alternative algorithms – Covariance models

- It is common to find examples of homologous RNAs that have a similar secondary structure without sharing significant sequence similarity.
- Areas of base pairing in tRNA can vary between various species. Sequence alignment can determine those areas.
- Base pairing creates the **same stable tRNA structure in different organisms**.
- But one mutation (variation) in one base can break down the structure.



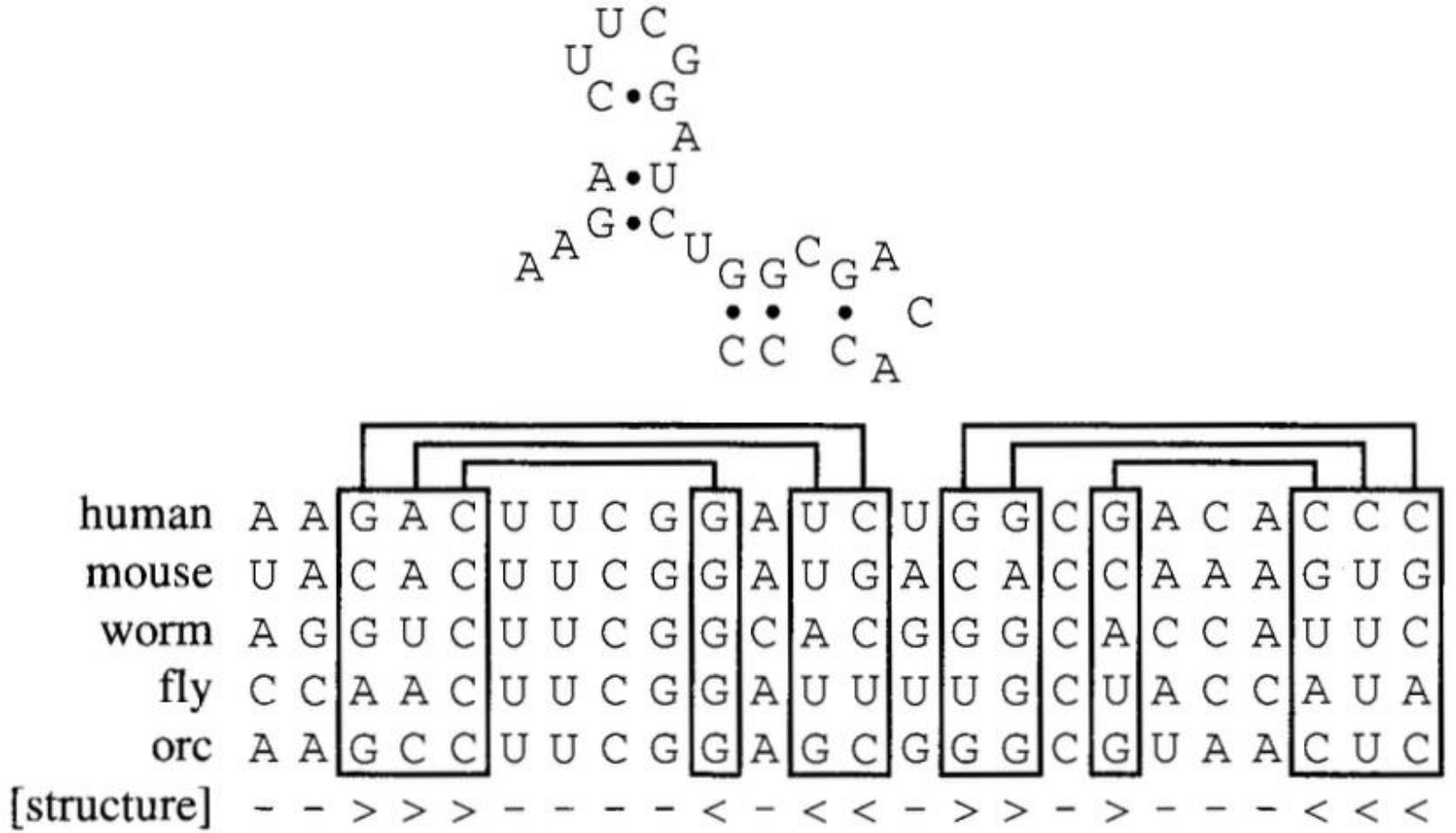
Source: *Bioinformatics*, Mount.

Identification of base covariation

This method examines aligned sequences of the same class of RNA molecules from different species:

- On the one hand, sequence change (mutations) should maintain the base pairing for secondary structure formation to conserve double-stranded regions in RNA molecules.
- On the other hand, sequence change in loops and single-stranded regions should not have such constraints.
- Joint substitutions or covariations are found to occur during the evolution of such genes.

Identification of base covariation



Base paired positions are boxed

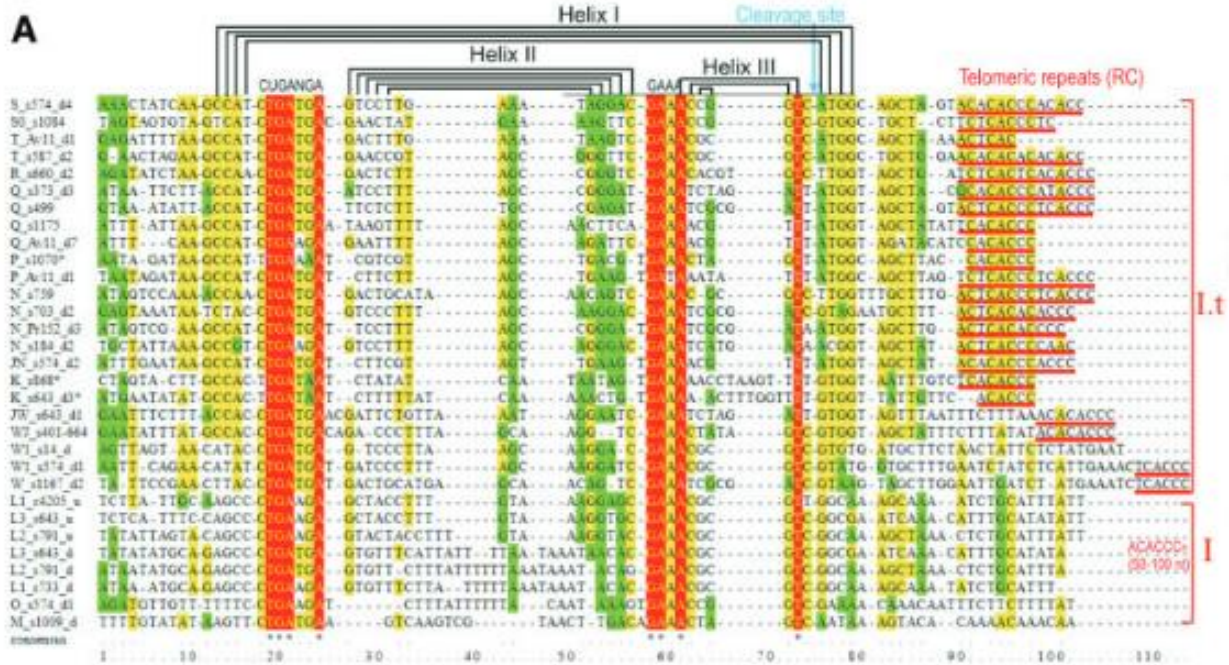
Consensus structure of an example RNA family (no gaps).

Source: *Biological sequence analysis*, Durbin *et al.*

Giant Reverse Transcriptase-Encoding Transposable Elements at Telomeres

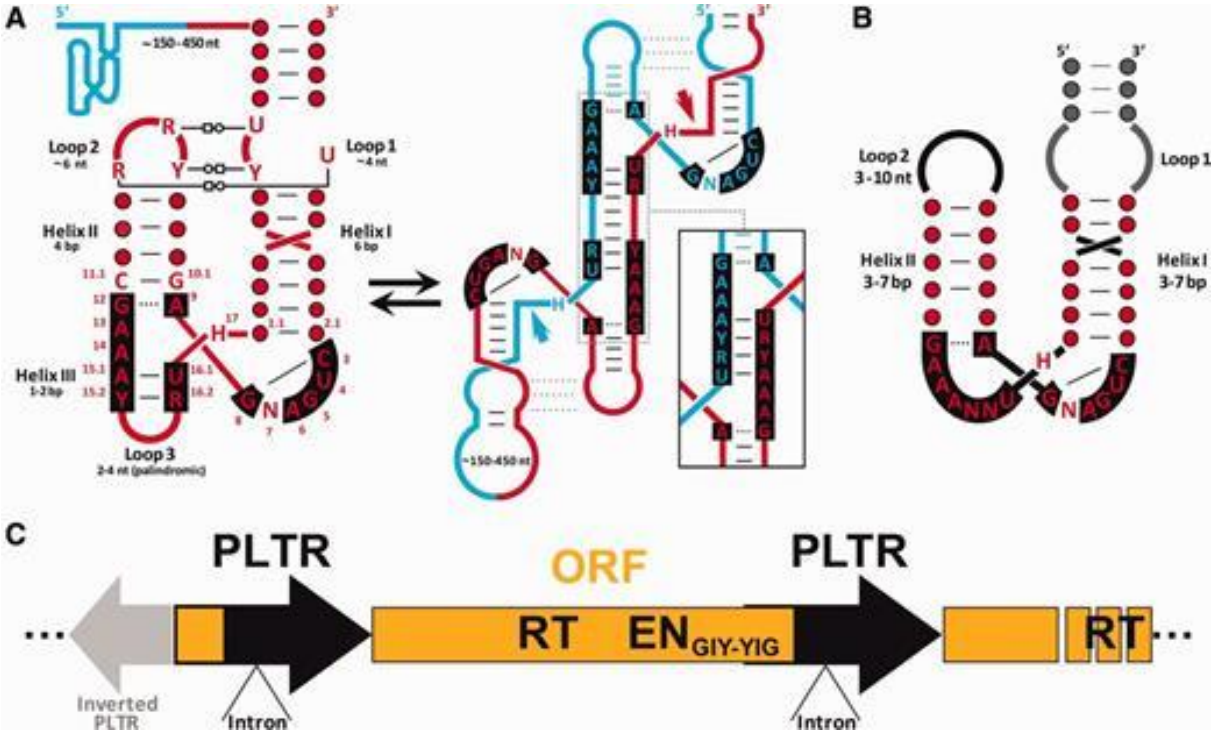
Irina R. Arkhipova,*[,1] Irina A. Yushenova,[1] and Fernando Rodriguez[1]
[1]Marine Biological Laboratory, Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Woods Hole, MA

link



Properties of PLE-associated **hammerhead ribozyme motif**

Eukaryotic *Penelope*-Like Retroelements Encode Hammerhead Ribozyme Motifs



HHR conformations in a tandem repeat, either as monomers (left) or as a dimeric active conformation (right).

Source: Cervera & De la Pena 2014. Link

RNAmotif Descriptor

```
parms
wc += gu;
descr
# descriptor for a min-HHR
# extra 5' sequence for later manual analysis
ss (len=20)
# stem I
h5 (minlen=3, maxlen=7)
# 1st catalytic center plus a possible extra N10
ss (minlen=7, maxlen=8, seq="^cuganga")
# stem II
h5 (minlen=3, maxlen=7)
# loop 2
ss (minlen=3, maxlen=10)
# stem II
h3
# 2nd catalytic center plus a possible extra N11
ss (minlen=8, maxlen=9, seq="gaaannuh$")
# stem I
h3
# extra 3' sequence for later manual analysis
ss (len=20)
```

Thursday Lab09

!

- **JupyterLab / Notebook: Lab09_Nussinov.ipynb (Content>Labs)**

https://jupyter.org/
https://jupyter.org/try-jupyter/lab/
https://code.visualstudio.com/

- ***RNAMotif***:

https://github.com/dacase/rnamotif

In *Oedipus*:
/mnt/sde_dir/software/rnamotif/rnamotif