Lecture 4: RNA folding

Chapter 6 – Problem 6.51 in Jones and Pevzner
and the *Turner model*

Spring 2020
February 4, 2020

RNA Basics

- RNA bases A,C,G,U
- Canonical Base Pairs
  - A-U
  - G-C
  - G-U “wobble” pairing
  - Bases can only pair with **one** other base.

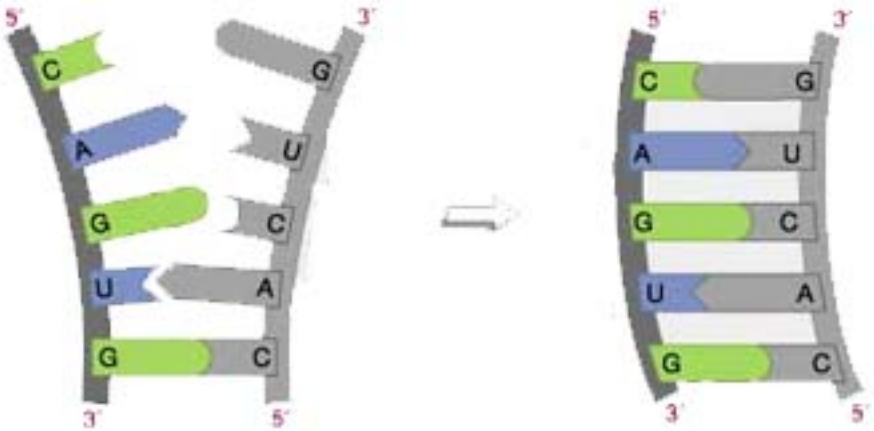
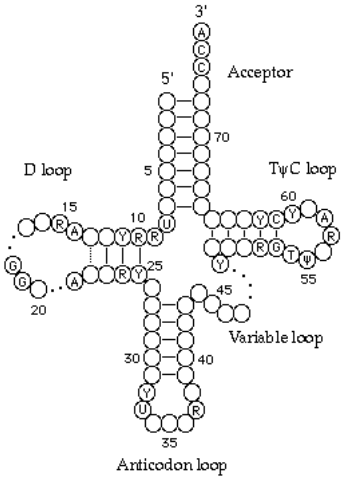


Image: http://www.bioalgorithms.info/

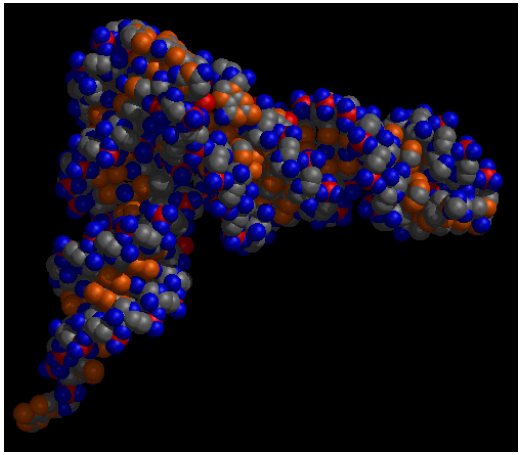
RNA Structural Levels

AAUCG…CUUCUUCCA

Primary

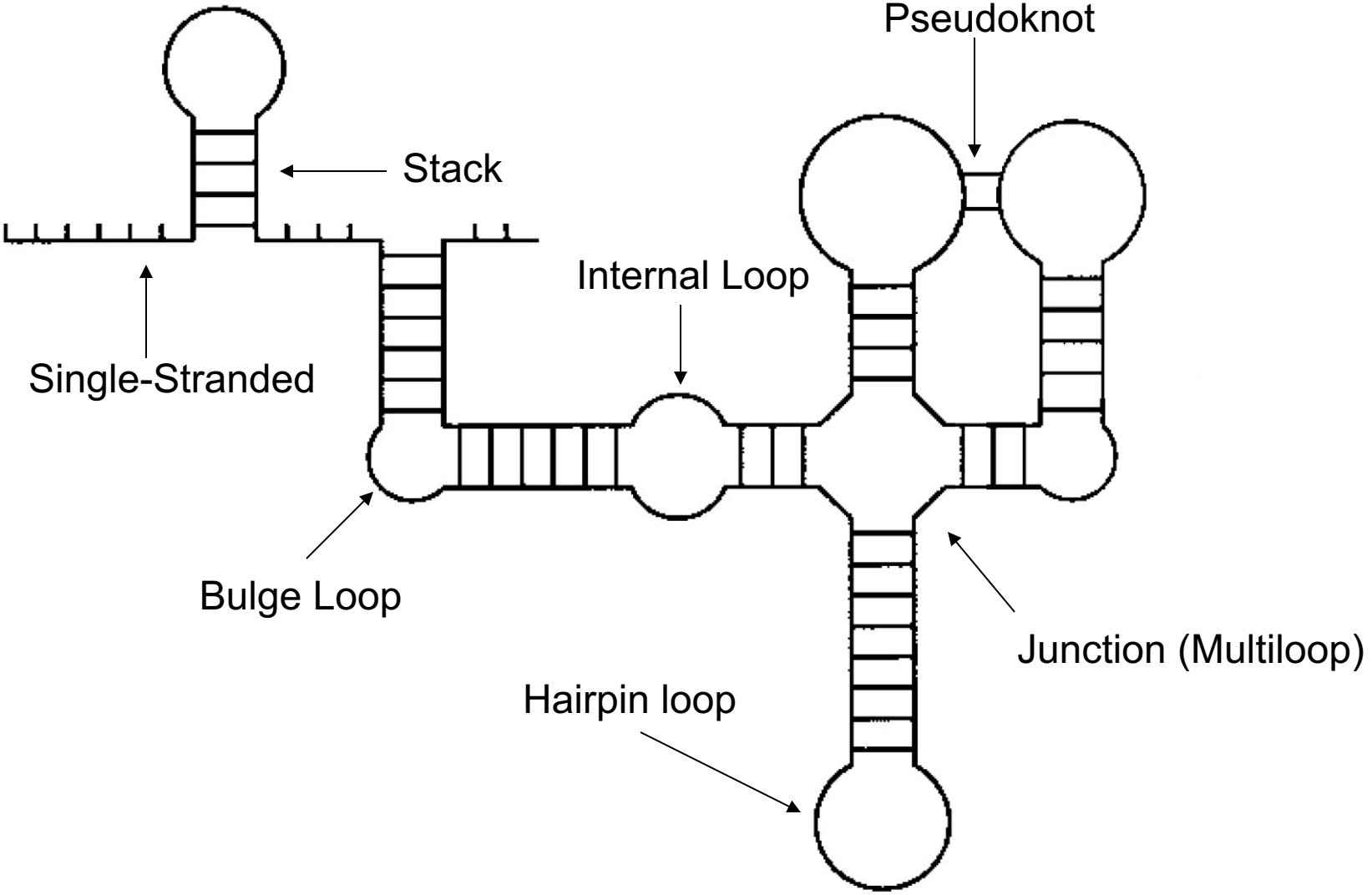


Secondary

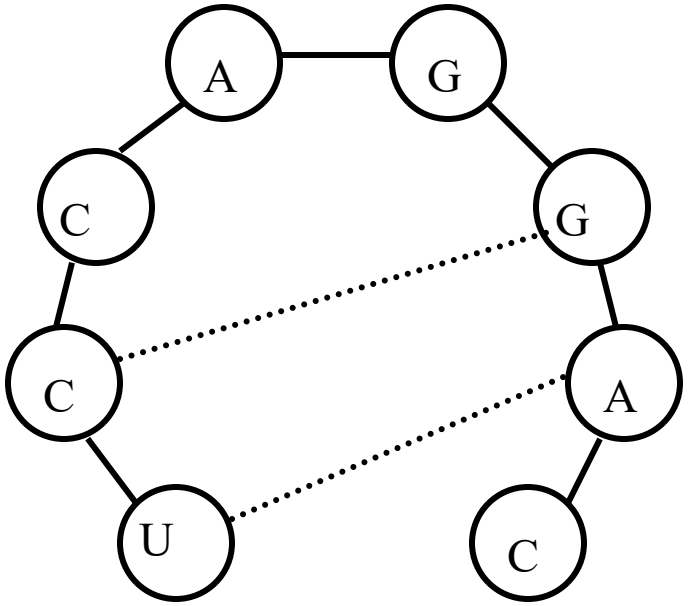


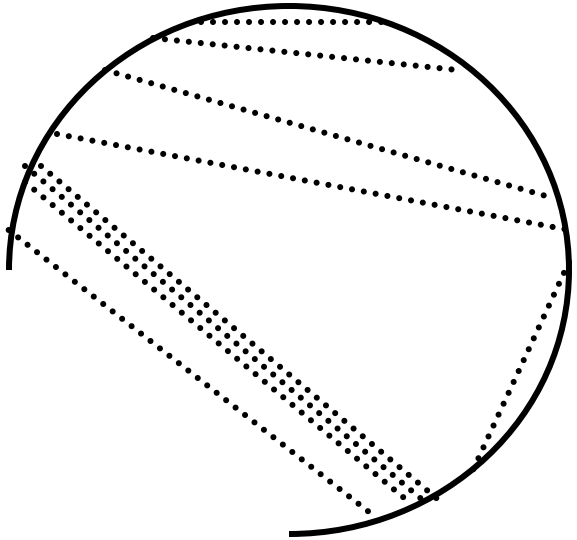
Tertiary

RNA Secondary Structure



Base Pair Maximization





Zuker (1981) *Nucleic Acids Research* 9(1) 133-149

Base Pair Maximization – Dynamic Programming Algorithm

Simple Example:
Maximizing Base Pairing

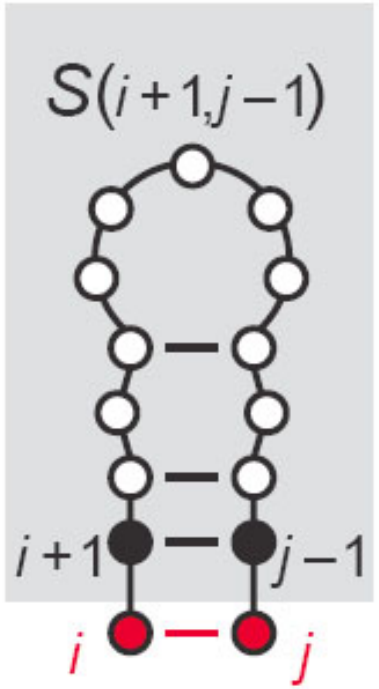
Base Pair Maximization – Dynamic Programming Algorithm

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

$$S(i,j) = \max \begin{cases} S(i+1,j-1)+1 \quad [\text{if } i,j \text{ base pair}] \\ S(i+1,j) \\ S(i,j-1) \\ \max_{i<k<j} S(i,k) + S(k+1,j) \end{cases}$$

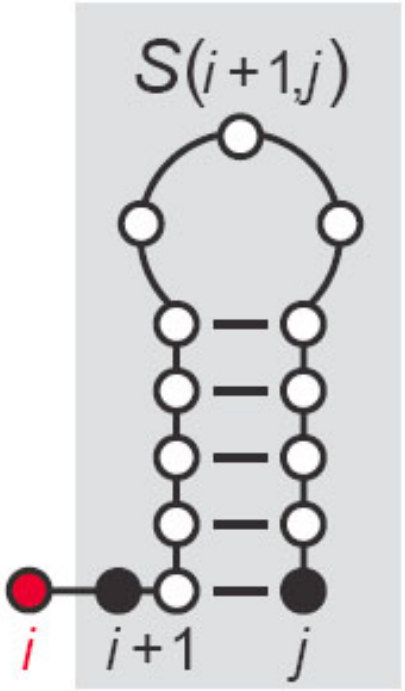
Base Pair Maximization – Dynamic Programming Algorithm

$$S(i,j) = \max \begin{cases} S(i+1,j-1) + 1 & [\text{if } i,j \text{ base pair}] \\ S(i+1,j) \\ S(i,j-1) \\ \max_{i<k<j} S(i,k) + S(k+1,j) \end{cases}$$



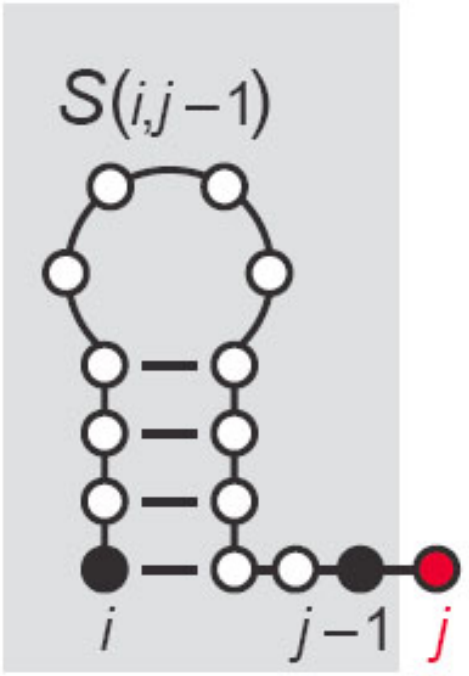
Base Pair Maximization – Dynamic Programming Algorithm

$$S(i,j) = \max \begin{cases} S(i+1,j-1)+1 & [\text{if } i,j \text{ base pair}] \\ S(i+1,j) \\ S(i,j-1) \\ \max_{i<k<j} S(i,k) + S(k+1,j) \end{cases}$$



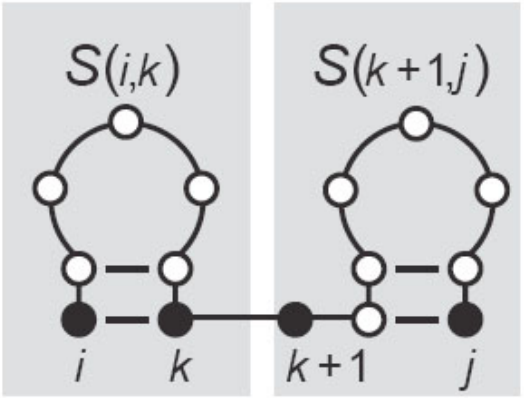
Base Pair Maximization – Dynamic Programming Algorithm

$$S(i,j) = \max \begin{cases} S(i+1,j-1)+1 \quad [\text{if } i,j \text{ base pair}] \\ S(i+1,j) \\ S(i,j-1) \\ \max_{i<k<j} S(i,k) + S(k+1,j) \end{cases}$$

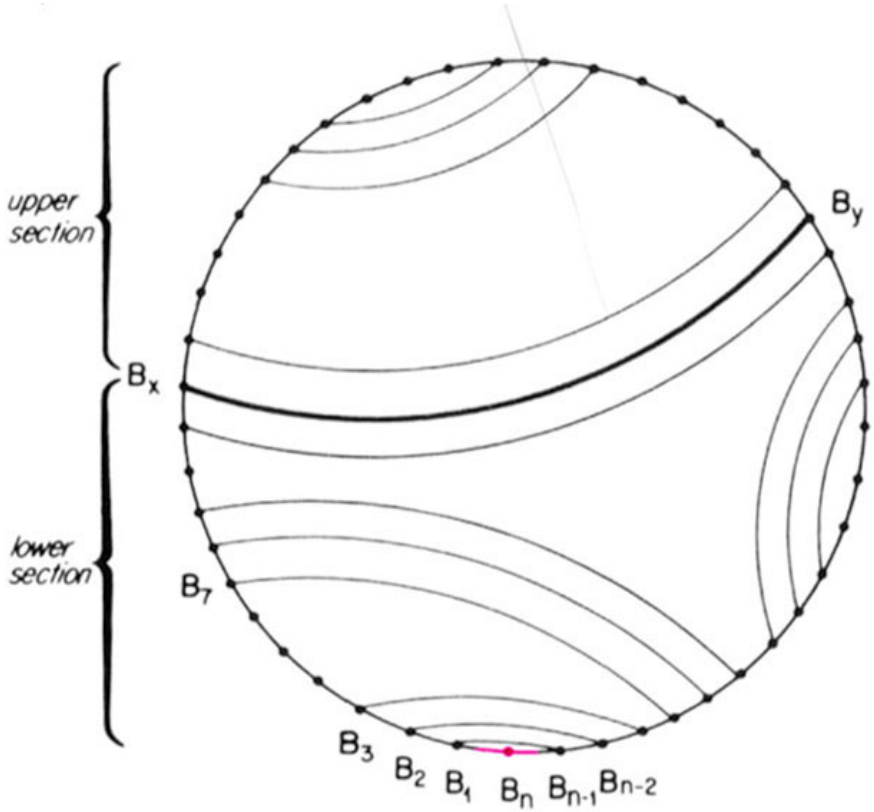


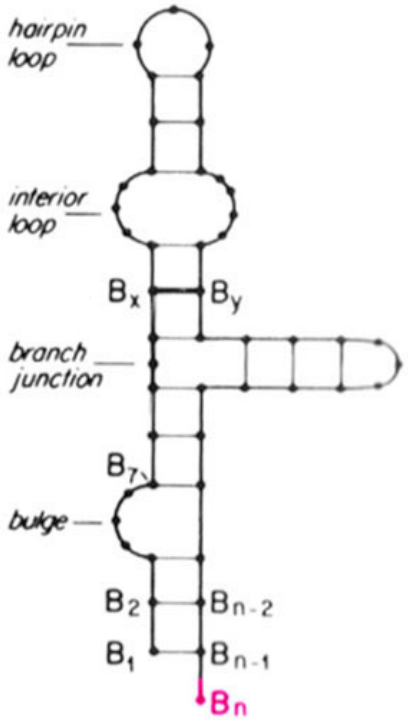
Base Pair Maximization – Dynamic Programming Algorithm

$$S(i,j) = \max \begin{cases} S(i+1,j-1)+1 & [\text{if } i,j \text{ base pair}] \\ S(i+1,j) \\ S(i,j-1) \\ \max_{i<k<j} S(i,k) + S(k+1,j) \end{cases}$$



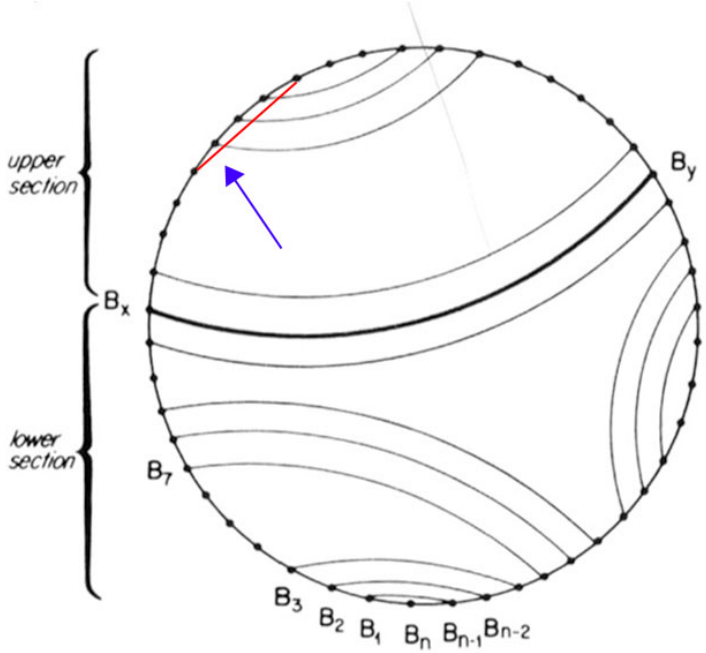
Circular Representation

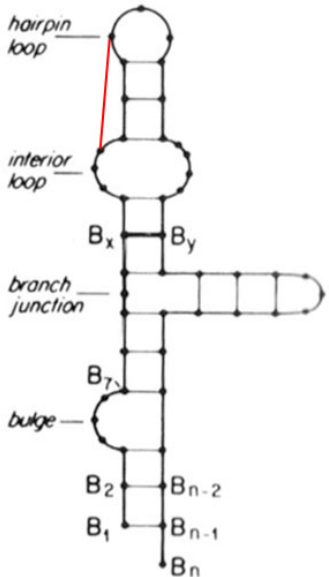




Images – David Mount

Pseudoknots





Images – David Mount

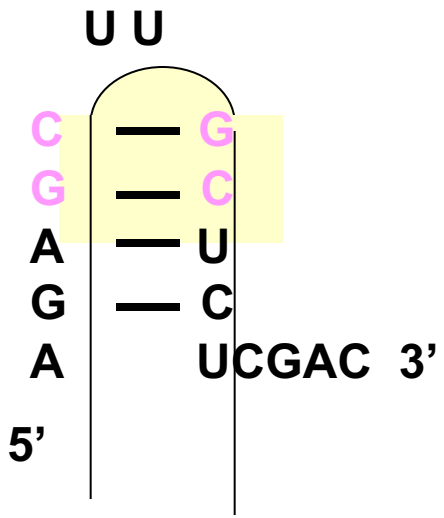
- Pseudoknots cause a breakdown in the presented Dynamic Programming Algorithm.
- In order to form a pseudoknot, checks must be made to ensure base is not already paired – this breaks down the divide and conquer recurrence relations.

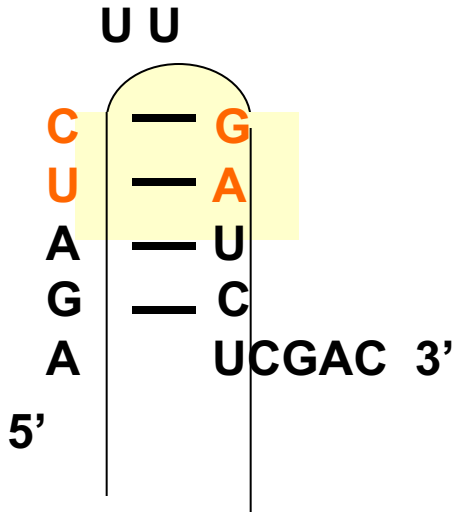
Simplifying Assumptions

- RNA folds into one minimum free-energy structure.
- There are no knots (base pairs never cross).
- The energy of a particular base pair in a double stranded region is sequence independent.
  - Neighbors do not influence the energy.

- Was solved by dynamic programming, Zucker and Steigler 1981

Sequence Dependent Base Pair Energy Values (Nearest Neighbor Model)

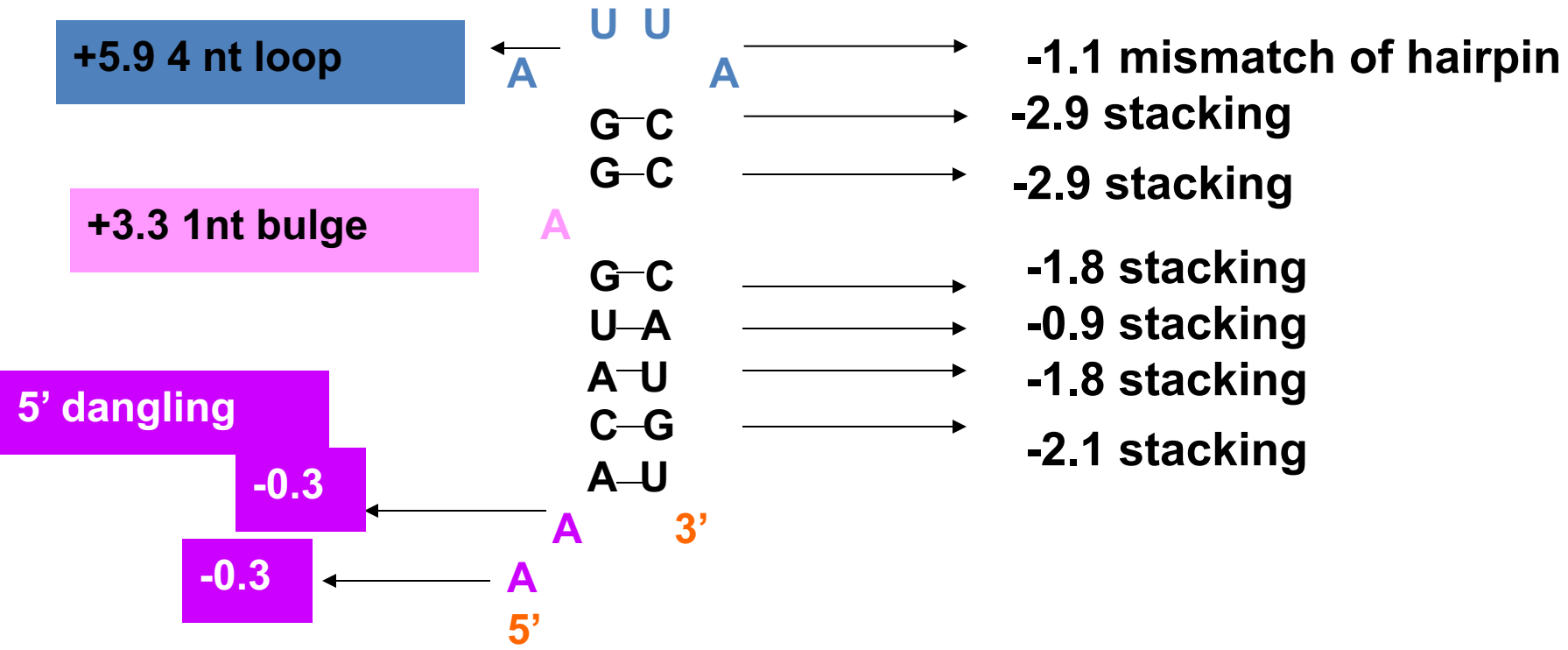




Example values:

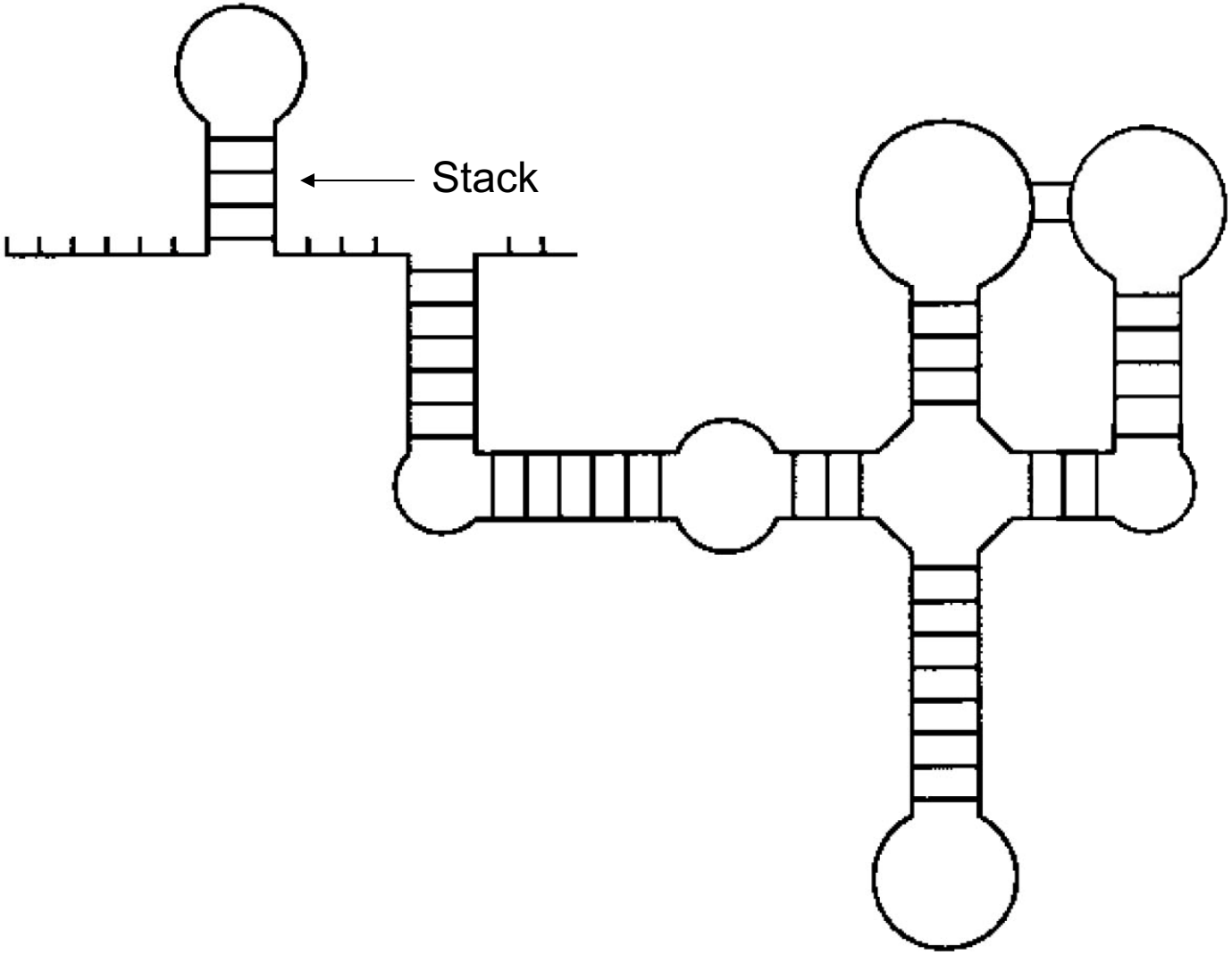
GC	GC	GC	GC
AU	GC	CG	UA
-2.3	-2.9	-3.4	-2.1

Free Energy Computation (Nearest Neighbor Model)



ΔG= - 4.9 kcal/mol

RNA Secondary Structure

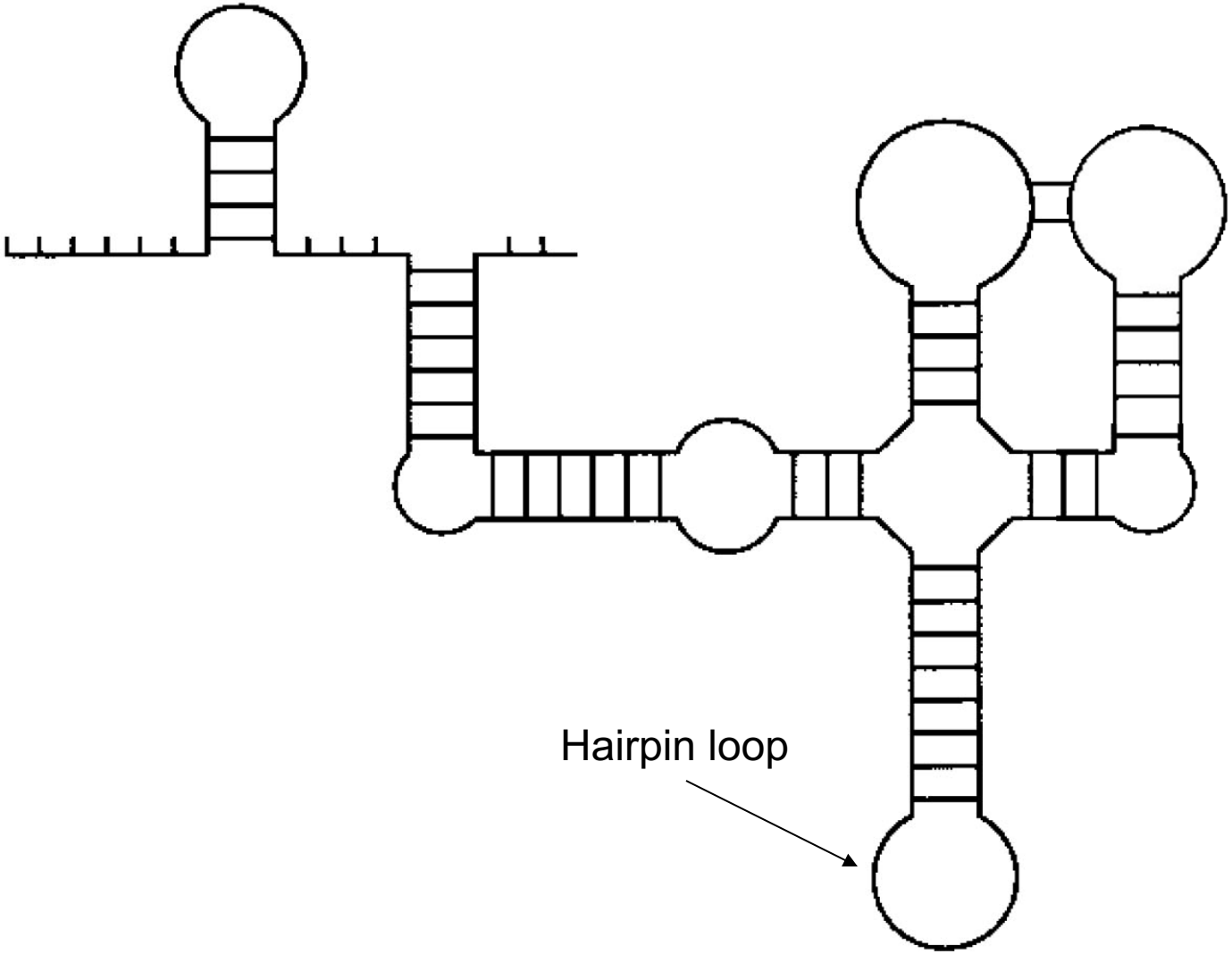


Nearest Neighbor Model

- Stacking energy - assign negative energies to these *between base pair* regions.
  - Energy is influenced by the nearest closing base pair
  - These energies are estimated experimentally from small synthetic RNAs.

- Positive energy - added for low entropy regions such as bulges, loops, etc.

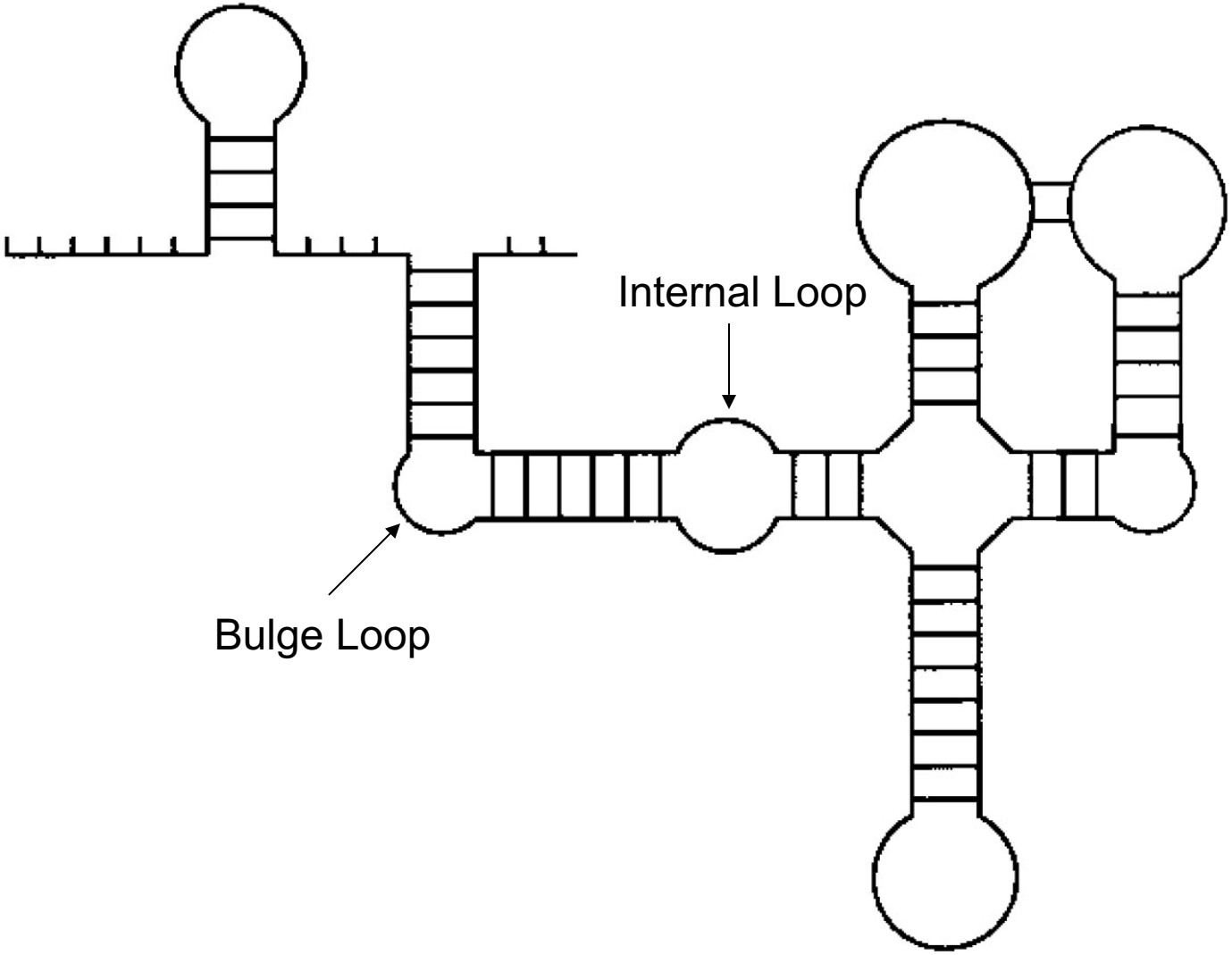
RNA Secondary Structure



Nearest Neighbor Model

- Hairpin energy:
  - Experimentally measured for hairpins of length 5, 6, 7, 8, … up to a maximum. Extrapolation above the maximum.
  - The closing pair affects the energy. Distinguish between A-U and C-G.

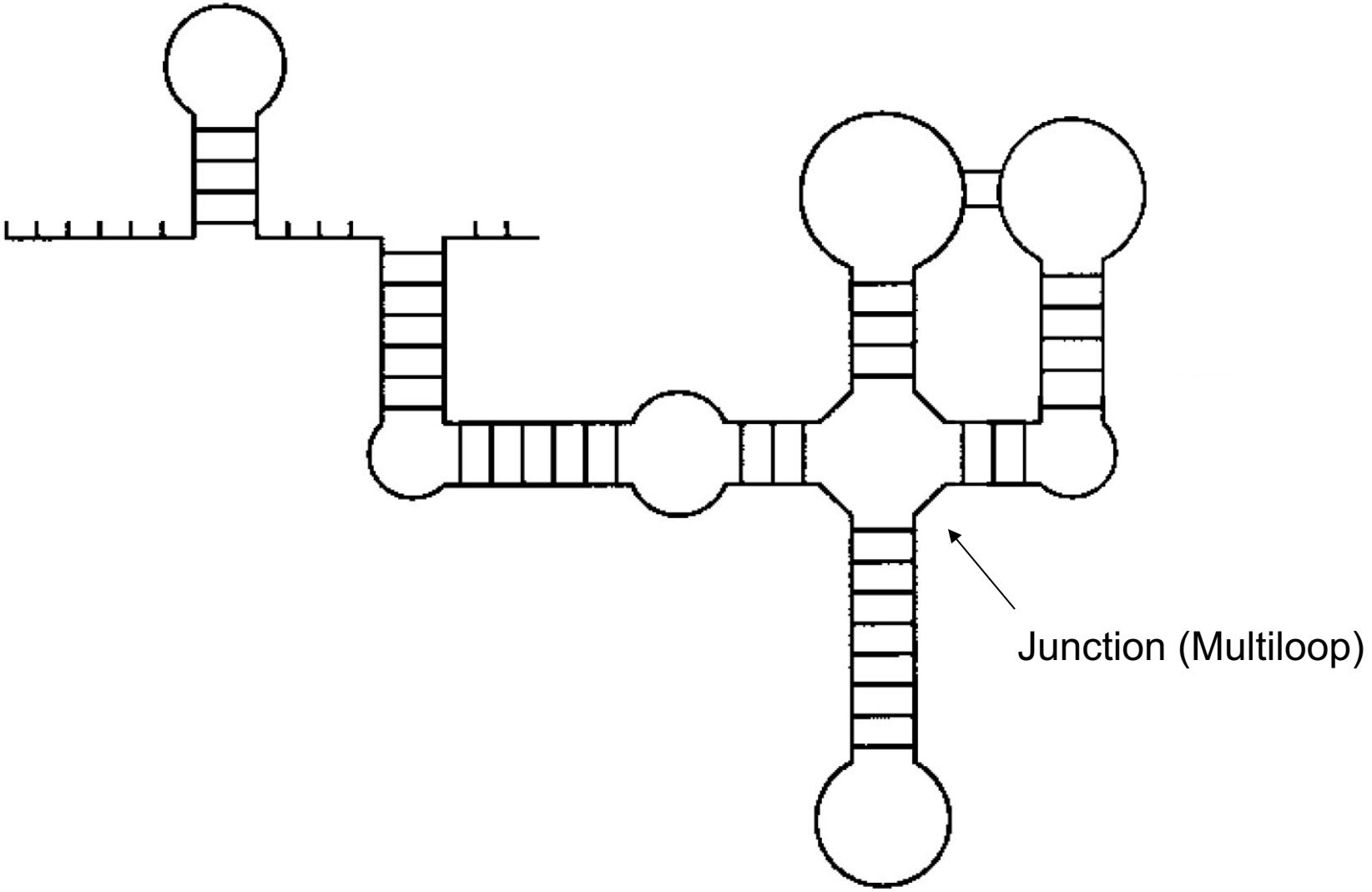
RNA Secondary Structure



Nearest Neighbor Model

- Bulge/Internal energy:
  - Let L_1, L_2 denote the lengths of the two sides of the bulge/internal loop.
  - Experimentally measured for different values of L_1, L_2.
  - In practice for computational convenience, the energy is given as function of $L_1 + L_2$ by a lookup table and extrapolation.

RNA Secondary Structure



Nearest Neighbor Model

- Multiloop energy:
  - Let U denote the number of unpaired bases.
  - Let P denote the number of base pairs.
  - The free energy is an affine function of U and P:

$$a_1 + a_2 U + a_3 P.$$

  - This is the least accurate component of the NN model.