

RNA Secondary Structure Prediction

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2020

Daifeng Wang

daifeng.wang@wisc.edu

Goals for Lecture

Key concepts

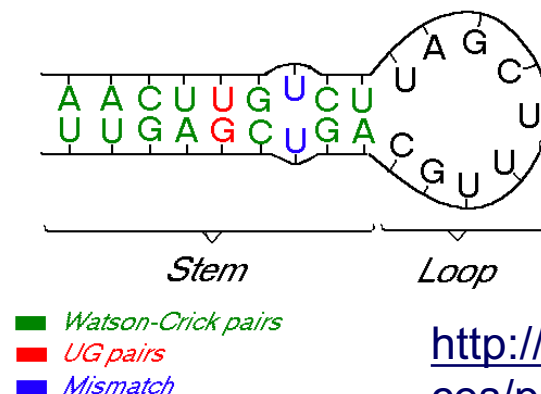
- RNA secondary structure
- Secondary structure features: stems, loops, bulges
- Pseudoknots
- Nussinov algorithm
- Adapting Nussinov to take free energy into account

Why RNA is Interesting

- Messenger RNA (mRNA) isn't the only important class of RNA
 - ribosomal RNA (rRNA)
 - ribosomes are complexes that incorporate several RNA subunits in addition to numerous protein units
 - transfer RNA (tRNA)
 - transport amino acids to the ribosome during translation
 - the spliceosome, which performs intron splicing, is a complex with several RNA units
 - microRNAs and others that play regulatory roles
 - many viruses (e.g. HIV, COVID-19) have RNA genomes
 - guide RNA
 - sequence complementary determines whether to cleave DNA
- Folding of an mRNA can be involved in regulating the gene's expression

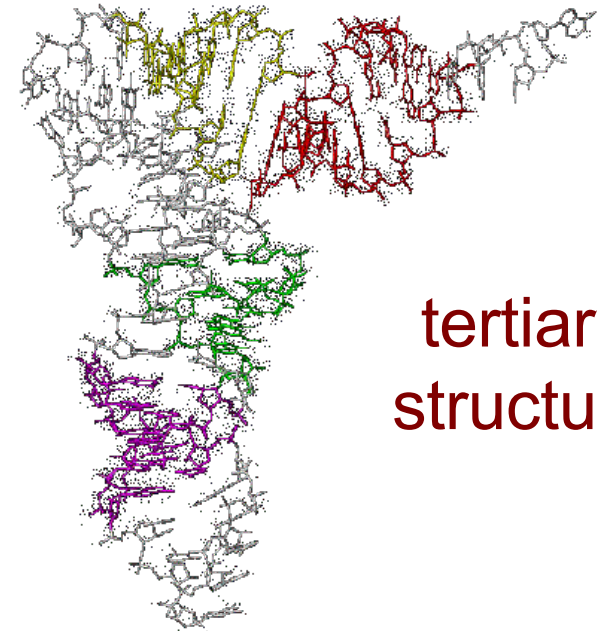
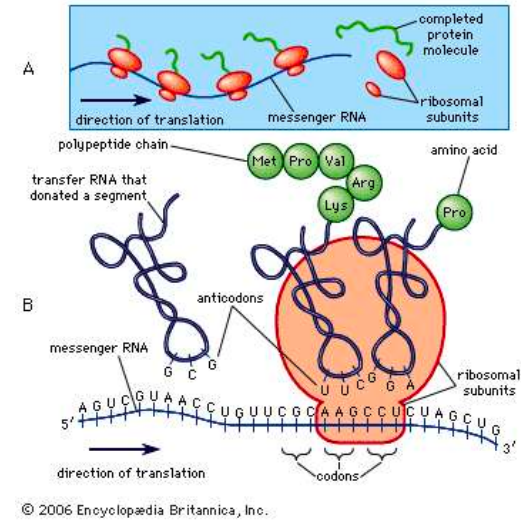
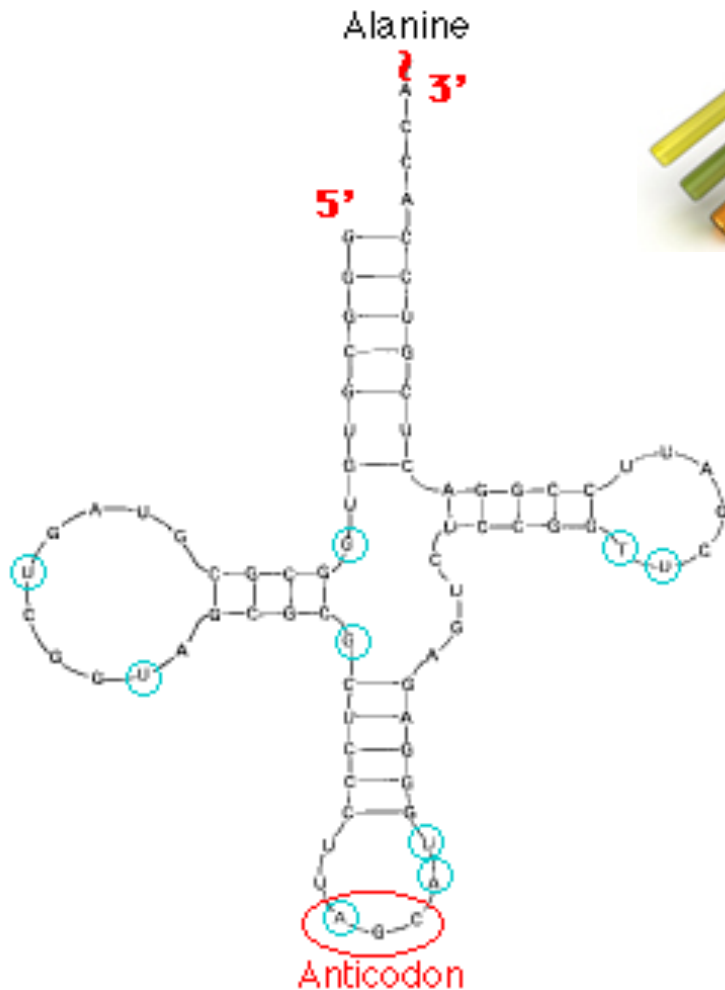
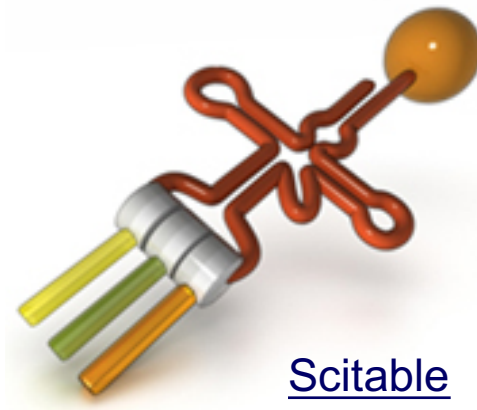
RNA Secondary Structure

- RNA is typically single stranded
- Folding, in large part is determined by base-pairing
 A–U and **C–G** are the canonical base pairs
 other bases will sometimes pair, especially **G–U**
- Base-paired structure is referred to as the *secondary structure* of RNA
- Related RNAs often have homologous secondary structure without significant sequence similarity

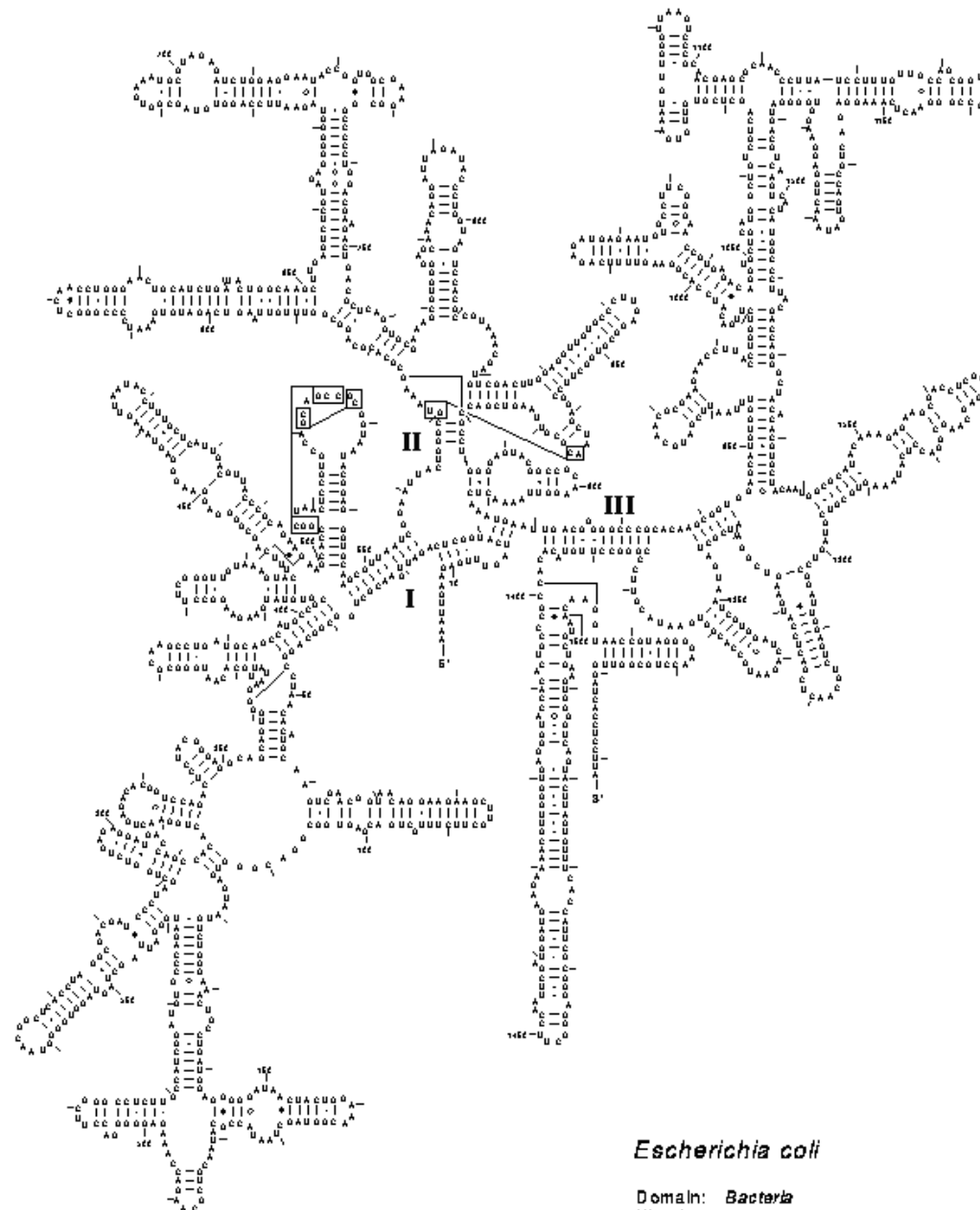


<http://www.bioinf.man.ac.uk/resources/phase/manual/node72.html>

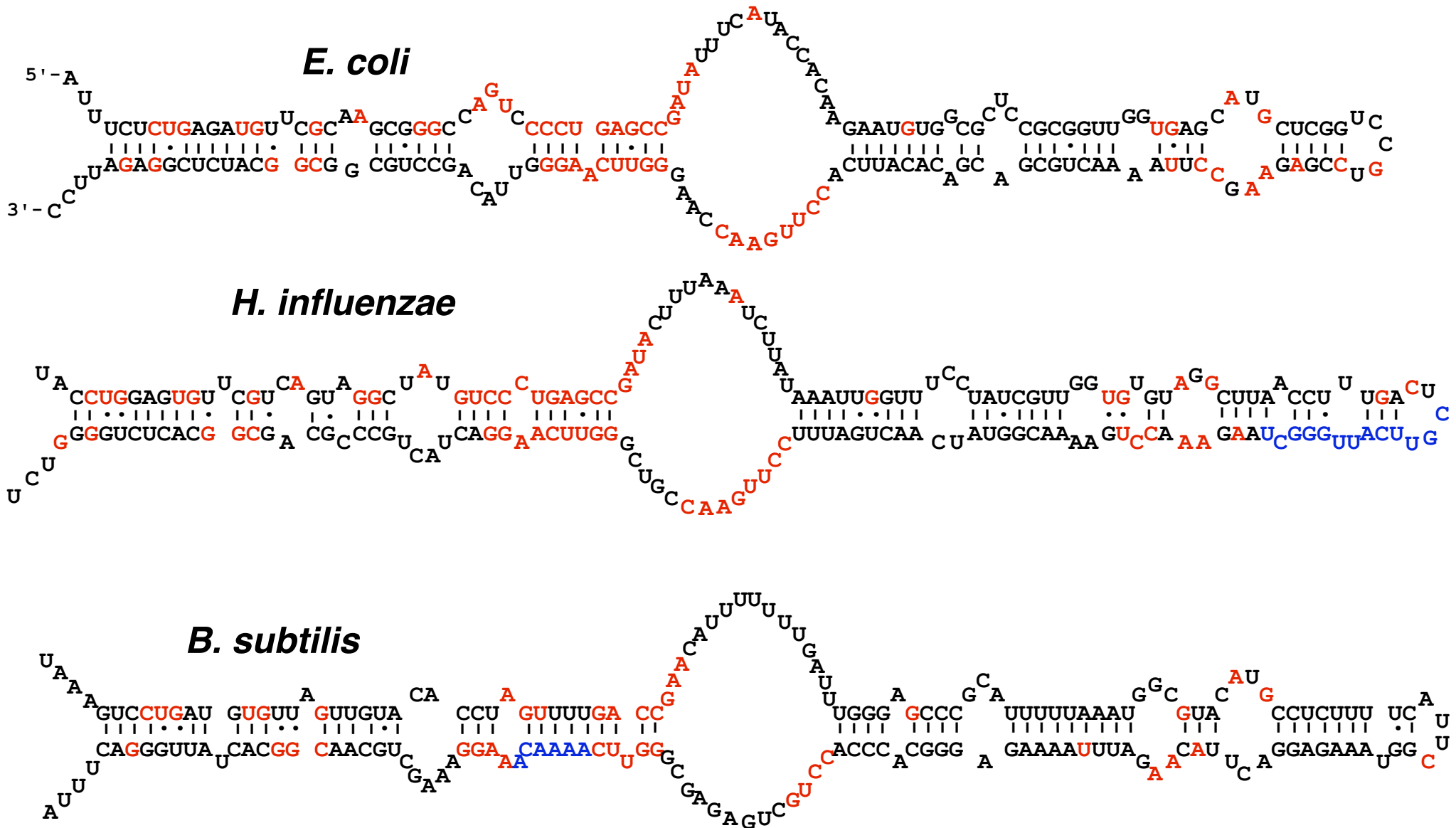
tRNA Secondary Structure



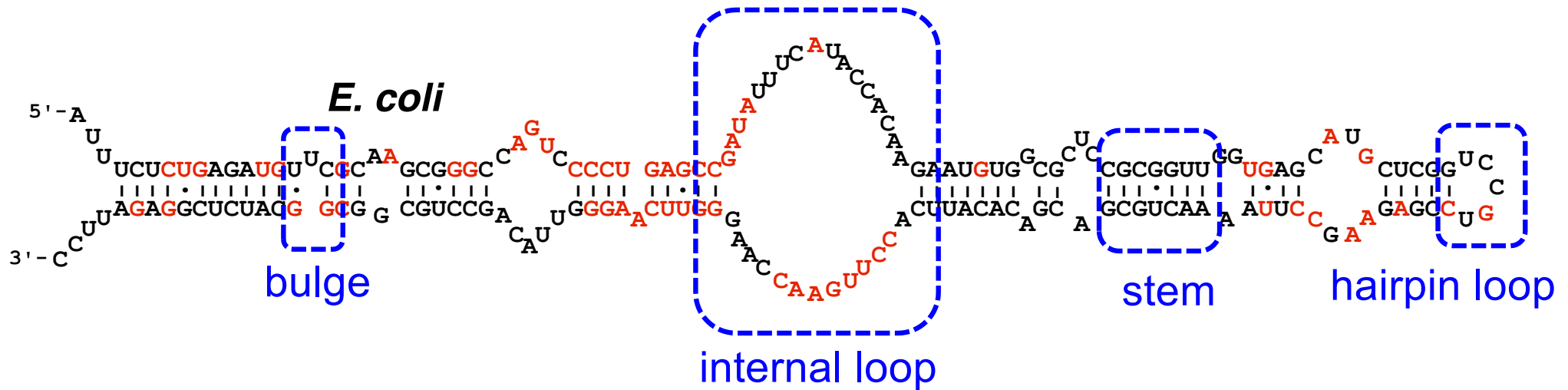
Small Subunit Ribosomal RNA



6S RNA Secondary Structure



Secondary Structure Features

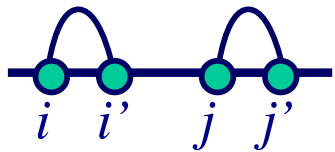


Four Key Problems

- Predicting RNA secondary structure Focus for today
Given: RNA sequence
Do: predict secondary structure that sequence will fold into
- Searching for instances of a given structure
Given: an RNA sequence or its secondary structure
Do: find sequences that will fold into a similar structure
- Modeling a family of RNAs
Given: a set of RNA sequences with similar secondary structure
Do: construct a model that captures the secondary structure regularities of the set
- Identifying novel RNA genes
Given: a pair of homologous DNA sequences
Do: identify subsequences that appear to have highly conserved RNA secondary structure (putative RNA genes)

RNA Folding Assumption

- Algorithms we'll consider assume that base pairings do not cross
- For base-paired positions i, i' and j, j' , with $i < i'$ and $j < j'$, we must have either

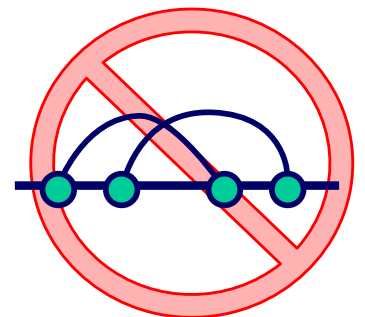


$i < i' < j < j'$ or $j < j' < i < i'$ (not nested)



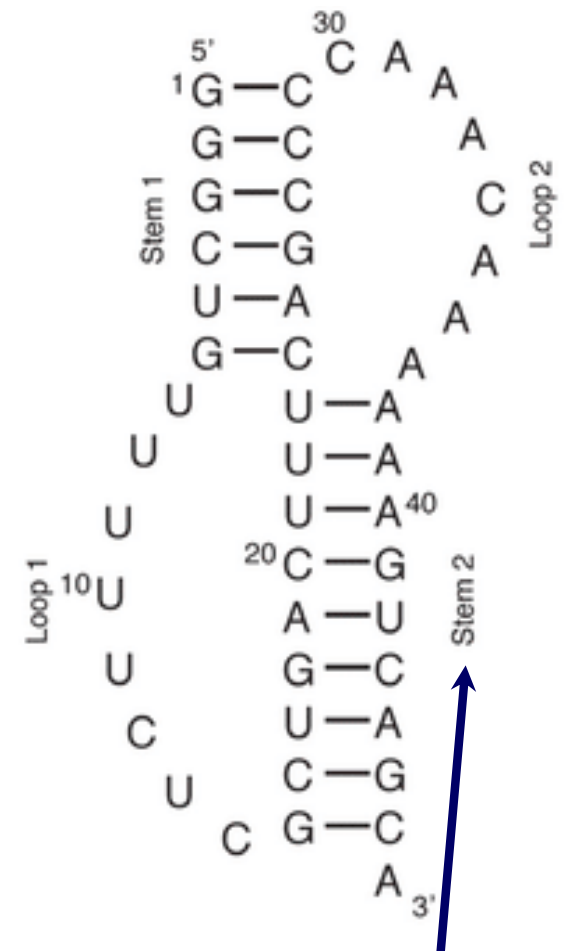
$i < j < j' < i'$ or $j < i < i' < j'$ (nested)

- Can't have $i < j < i' < j'$ or $j < i < j' < i'$



Pseudoknots

- These crossings are called *pseudoknots*
- Dynamic programming breaks down if pseudoknots are allowed
- Fortunately, they are not very frequent
- Modern software does support them
 - [Akiyama et al. 2018](#)



Pseudoknot between Stem 1 and 2

Simplest RNA Secondary Structure Task

Given:

- An RNA sequence
- The constraint that pseudoknots are not allowed

Do:

- Find a secondary structure for the RNA that maximizes the number of base pairing positions

Predicting RNA Secondary Structure: the Nussinov Algorithm

[Nussinov et al., *SIAM Journal of Applied Mathematics* 1978]

Key idea:

- Do this using dynamic programming
 - start with small subsequences
 - progressively work to larger ones

Idea of Dynamic Programming (DP)

- Given a problem, take advantage of the solutions for sub-problems and combine them to obtain the optimal solution
 - Recall sequence alignment: optimal aligning two sequences can be reduced to find optimal alignments of their all possible prefixes
 - Computational efficiency: reuse the alignment scores already calculated
- Model the problem via a recursive formula

DP in the Nussinov Algorithm

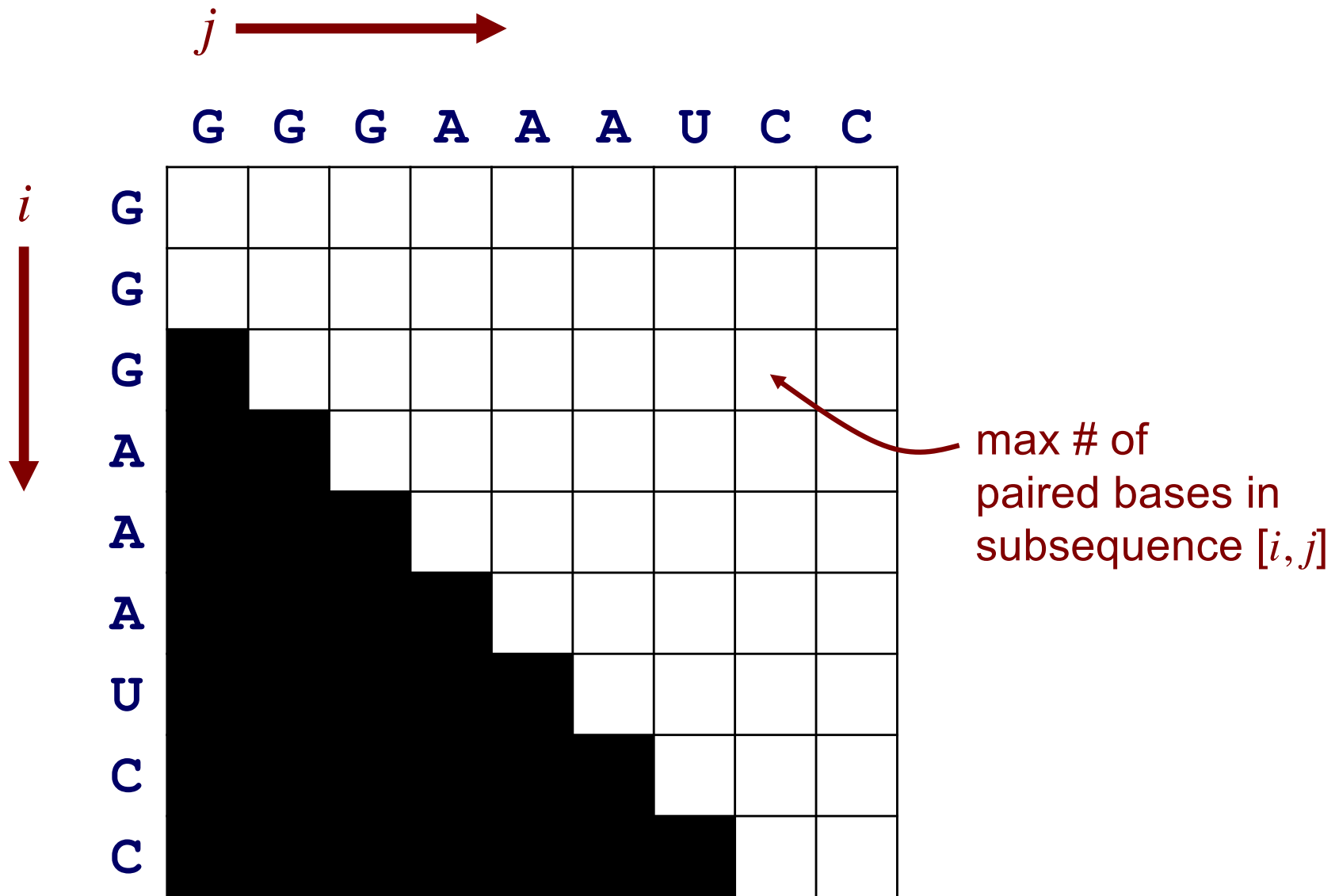


Figure 10.8 from textbook

DP in the Nussinov Algorithm

- Let $\delta(i, j) = \begin{cases} 1 & \text{if } x_i \text{ and } x_j \text{ are complementary} \\ 0 & \text{otherwise} \end{cases}$

- Initialization:

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

$$\gamma(i, i) = 0 \quad \text{for } i = 1 \text{ to } L$$

- Recursion

max # of
paired bases in
subsequence $[i, j]$


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

Nussinov Algorithm Traceback

push $(1, L)$ onto stack

repeat until stack is empty

pop (i, j)

if $i \geq j$ continue

else if $\gamma(i+1, j) = \gamma(i, j)$ push $(i+1, j)$

else if $\gamma(i, j-1) = \gamma(i, j)$ push $(i, j-1)$

else if $\gamma(i+1, j-1) + \delta(i, j) = \gamma(i, j)$

record i, j base pair

push $(i+1, j-1)$

else for $k = i+1$ to $j-1$: if $\gamma(i, k) + \gamma(k+1, j) = \gamma(i, j)$

push $(k+1, j)$

push (i, k)

break

Predicting RNA Secondary Structure by Energy Minimization

- It's naïve to predict folding just by maximizing the number of base pairs
- However, we can generalize the key recurrence relation so that we're minimizing free energy instead

$$E(i, j) = \min \begin{cases} E(i+1, j) \\ E(i, j-1) \\ \min_{i < k < j} [E(i, k) + E(k+1, j)] \\ P(i, j) \end{cases}$$

← case that i and j are base paired

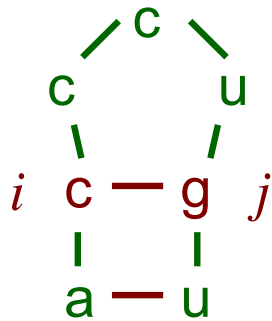
Predicting RNA Secondary Structure by Energy Minimization

- A sophisticated program, such as Mfold [Zuker et al.], can take into account free energy of the “local environment” of $[i, j]$

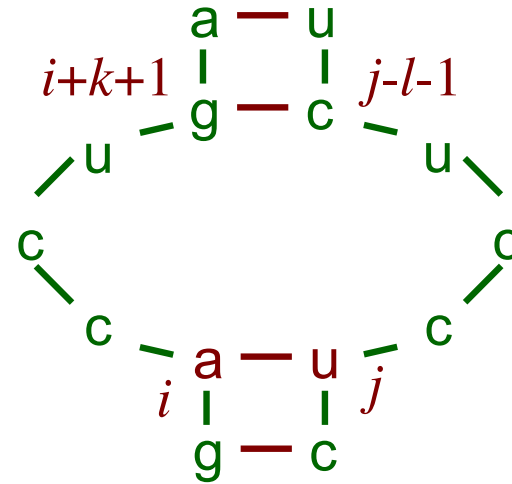
$$P(i, j) = \min \left\{ \begin{array}{l} \alpha(i, j) + \text{LoopEnergy}(j - i - 1) \\ \alpha(i, j) + \text{StackingEnergy}(i, j, i + 1, j - 1) + P(i + 1, j - 1) \\ \min_{k \geq 1} [\alpha(i, j) + \text{BulgeEnergy}(k) + P(i + k + 1, j - 1)] \\ \min_{k \geq 1} [\alpha(i, j) + \text{BulgeEnergy}(k) + P(i + 1, j - k - 1)] \\ \min_{k, l \geq 1} [\alpha(i, j) + \text{LoopEnergy}(k + l) + P(i + k + 1, j - l - 1)] \\ \min_{j > k > i} [\alpha(i, j) + E(i + 1, k) + E(k + 1, j - 1)] \end{array} \right.$$

Predicting RNA Secondary Structure by Energy Minimization

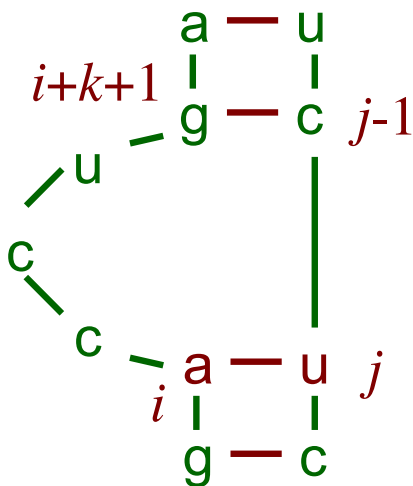
$$\alpha(i, j) + \text{LoopEnergy}(j - i - 1)$$



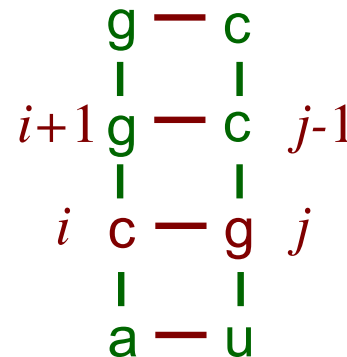
$$\min_{k, l \geq 1} [\alpha(i, j) + \text{LoopEnergy}(k + l) + P(i + k + 1, j - l - 1)]$$



$$\min_{k \geq 1} [\alpha(i, j) + \text{BulgeEnergy}(k) + P(i + k + 1, j - 1)]$$

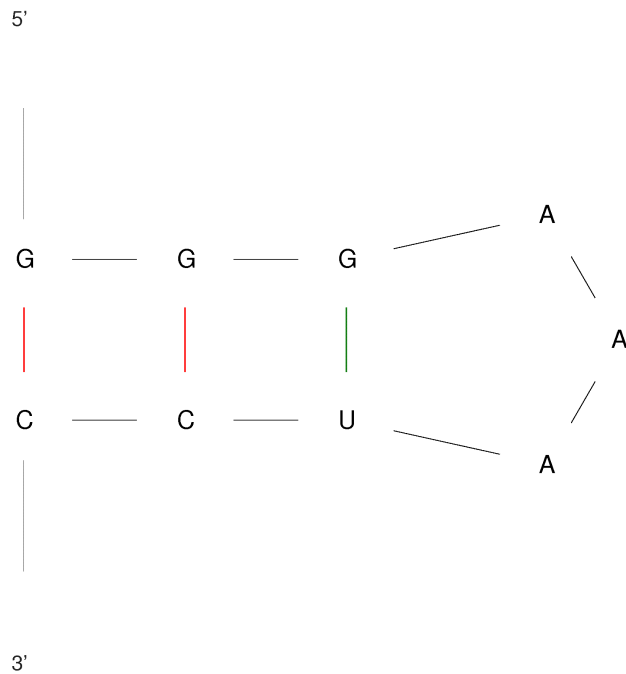


$$\alpha(i, j) + \text{StackingEnergy}(i, j, i + 1, j - 1) + P(i + 1, j - 1)$$

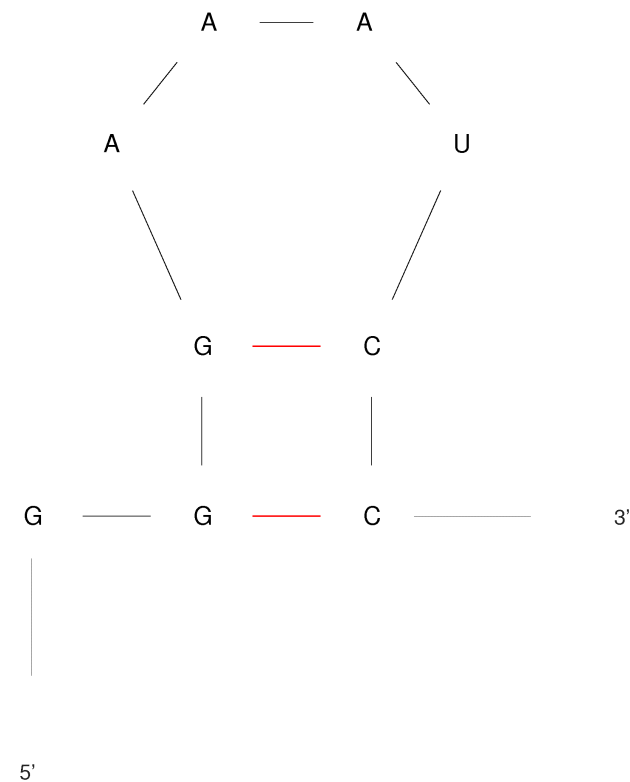


Mfold example

GGGAAUCC



$\Delta G = -0.80$ kcal/mol



$\Delta G = 0.20$ kcal/mol