# RNA Secondary Structure Prediction

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2017

Anthony Gitter

gitter@biostat.wisc.edu

These slides, excluding third-party material, are licensed under CC BY-NC 4.0 by Mark Craven, Colin Dewey, and Anthony Gitter

# 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) 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

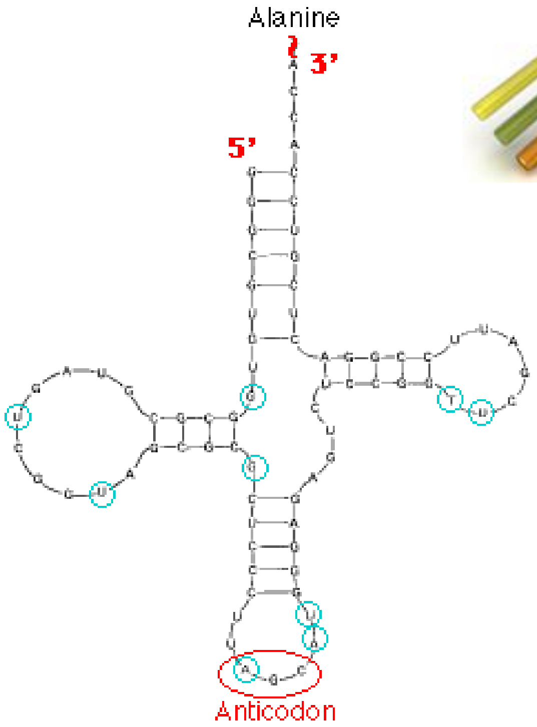
# RNA Secondary Structure

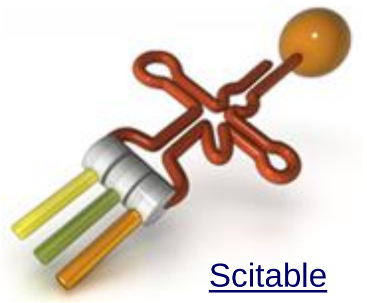
- RNA is typically single stranded
- Folding, in large part is determined by base-pairing

  **A–U** and **C–G** are the canonical base pairs

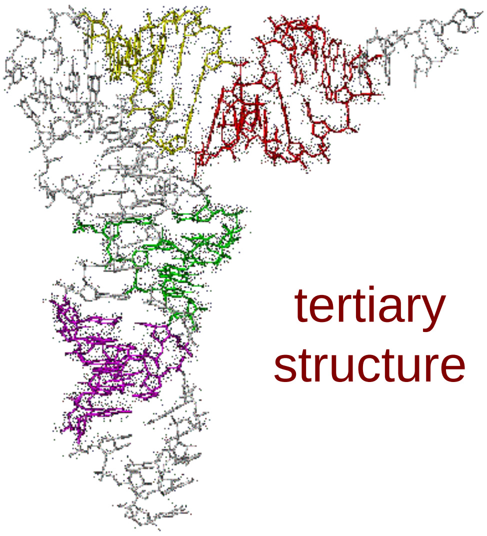
  other bases will sometimes pair, especially **G–U**
- Base-paired structure is referred to as the *secondary structure* of RNA
- Related RNAs often have homologous secondary structure without significant sequence similarity

# tRNA Secondary Structure



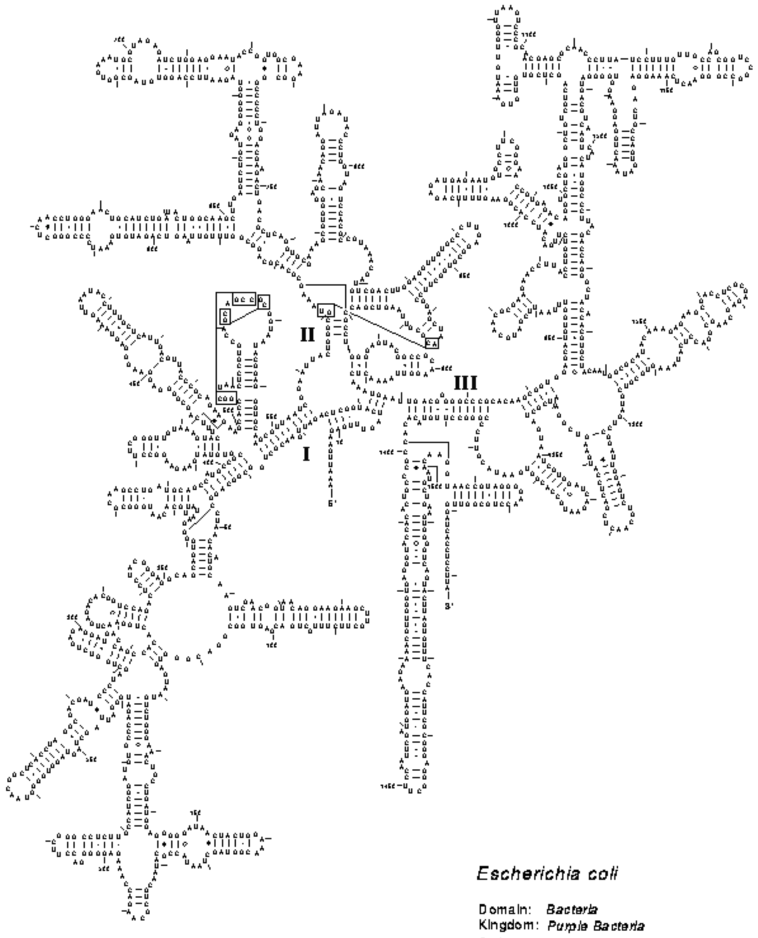


Scitable

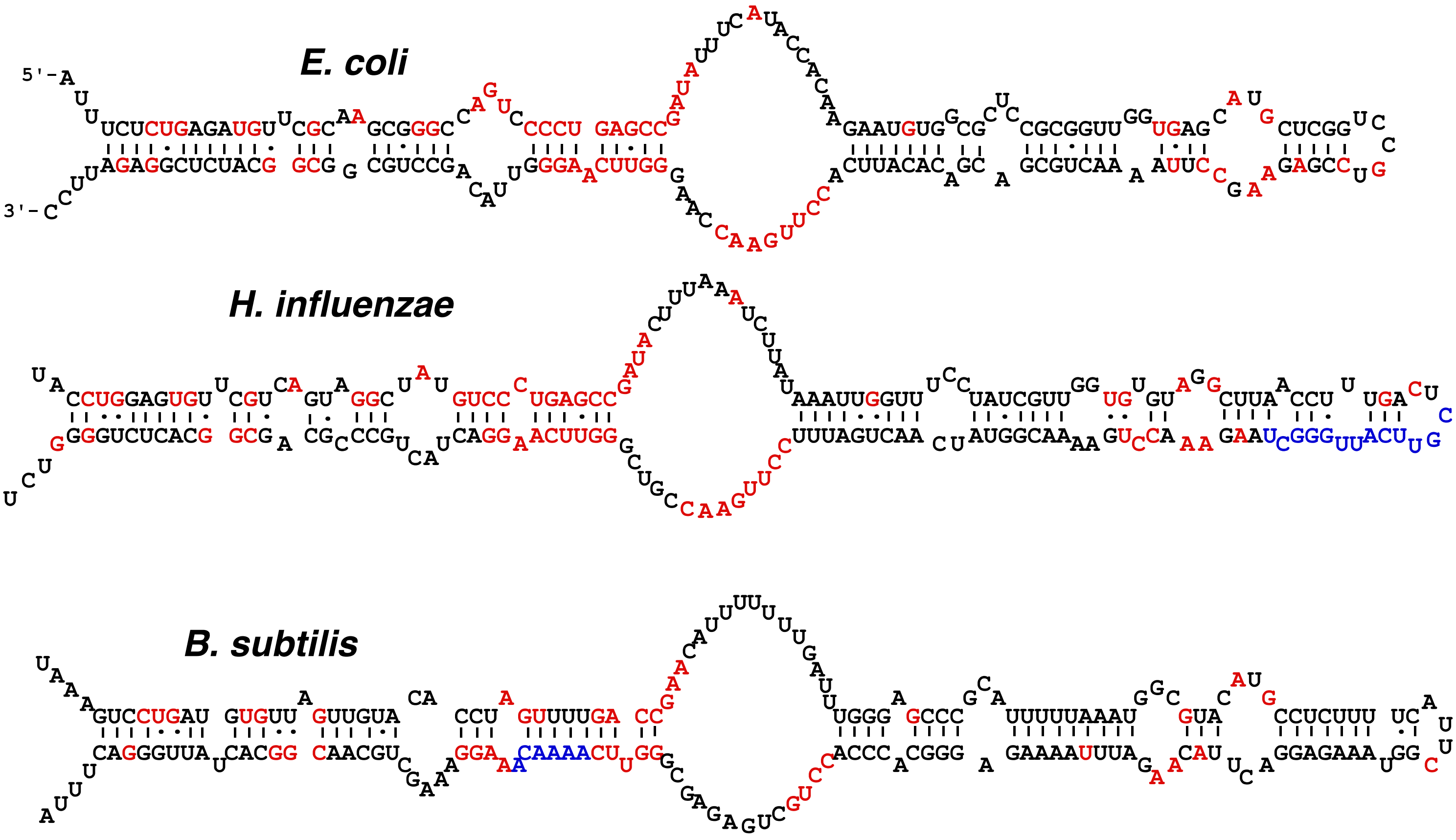


tertiary structure

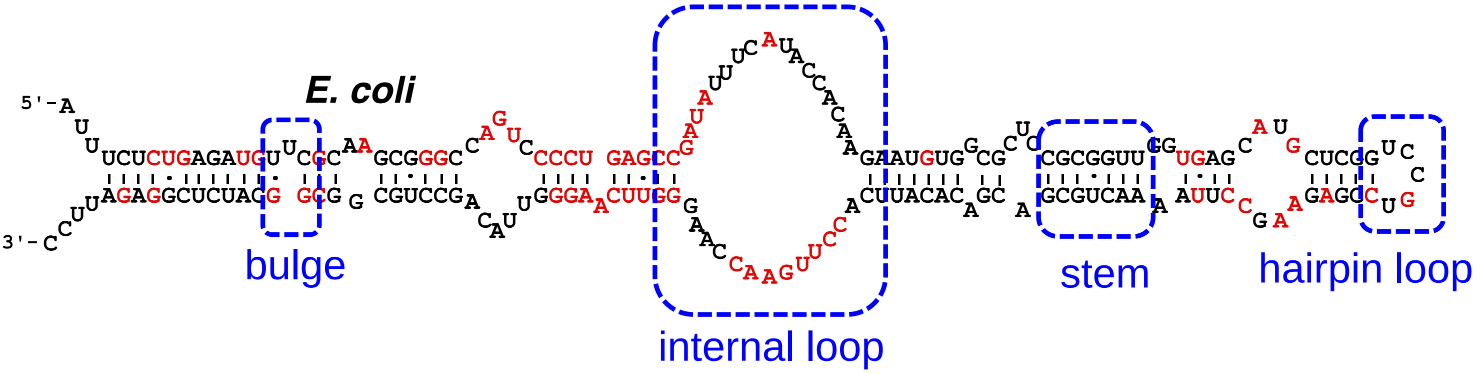
# Small Subunit Ribosomal RNA



# 6S RNA Secondary Structure

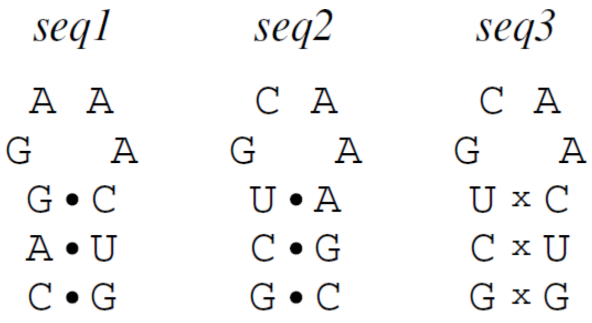


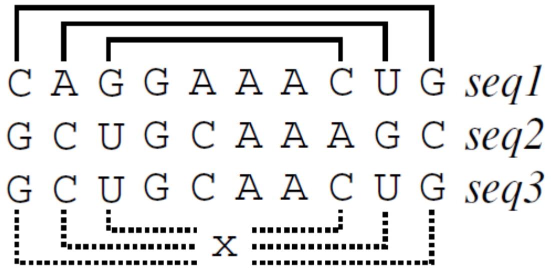
# Secondary Structure Features



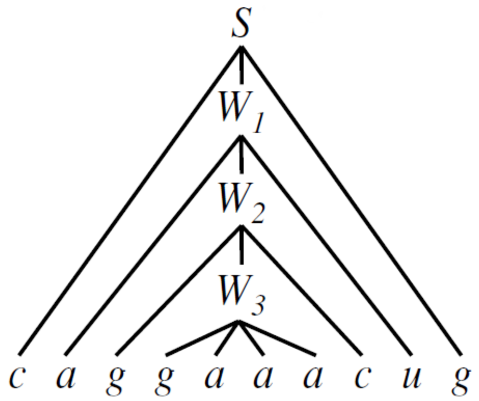
# Secondary structure as CFG

- Context-free grammar (CFG) is a suitable formalism for representing palindrome languages





$$
\begin{array}{lcl}
S & \rightarrow & aW_1u \mid cW_1g \mid gW_1c \mid uW_1a, \\
W_1 & \rightarrow & aW_2u \mid cW_2g \mid gW_2c \mid uW_2a, \\
W_2 & \rightarrow & aW_3u \mid cW_3g \mid gW_3c \mid uW_3a, \\
W_3 & \rightarrow & gaaa \mid gcaa.
\end{array}
$$



# Four Key Problems

- Predicting RNA secondary structure
  **Given**: RNA sequence
  **Do**: predict secondary structure that sequence will fold into

Focus for today

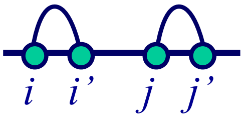
- Searching for instances of a given structure
  **Given**: an RNA sequence or its secondary structure
  **Do**: find sequences that will fold into a similar structure

- 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

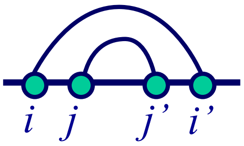
- Identifying novel RNA genes
  **Given**: a pair of homologous DNA sequences
  **Do**: identify subsequences that appear to have highly conserved RNA secondary structure (putative RNA genes)

# 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

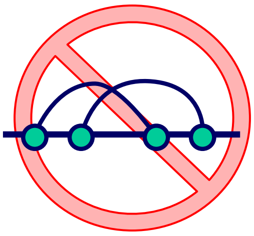


$i < i' < j < j'$ or $j < j' < i < i'$ (not nested)

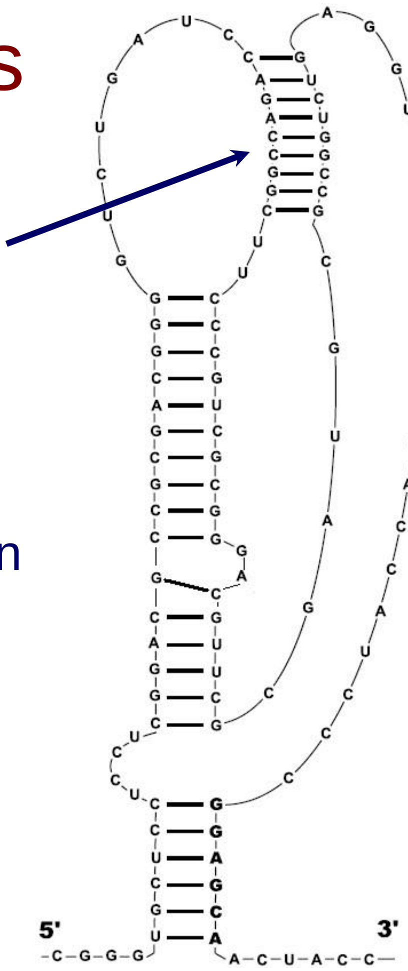


$i < j < j' < i'$ or $j < i < i' < j'$ (nested)

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



# Pseudoknots



pseudoknot

- These crossings are called *pseudoknots*
- Dynamic programming breaks down if pseudoknots are allowed
- Fortunately, they are not very frequent

Figure from Seliverstov et al. *BMC Microbiology*, 2005

# 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

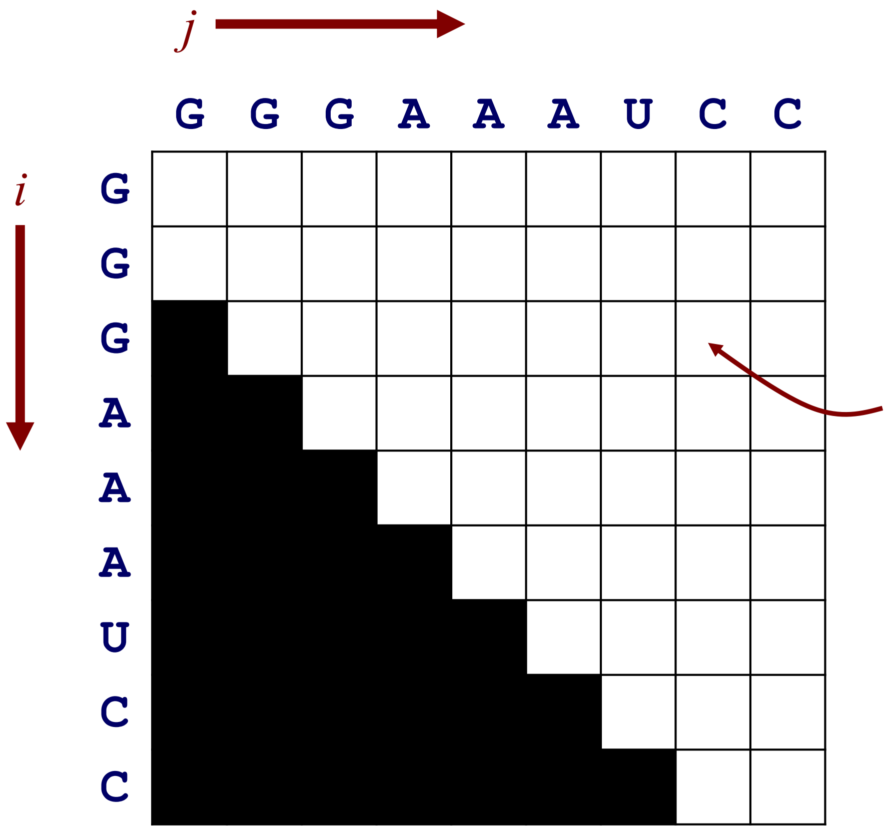
# Predicting RNA Secondary Structure: the Nussinov Algorithm

[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

# DP in the Nussinov Algorithm



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

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:

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

- Recursion

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

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

# 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 <u>minimizing</u> free energy instead

$$E(i, j) = \min \begin{cases} E(i+1, j) \\ E(i, j\text{-}1) \\ \min\limits_{\mathrm{i<k<j}} \left[ E(i,k) + E(k+1, j) \right] \\ P(i, j) \end{cases}$$

case that $i$ and $j$ are base paired

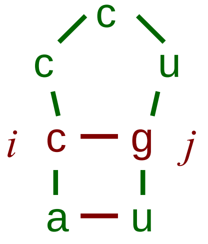
# Predicting RNA Secondary Structure by Energy Minimization

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

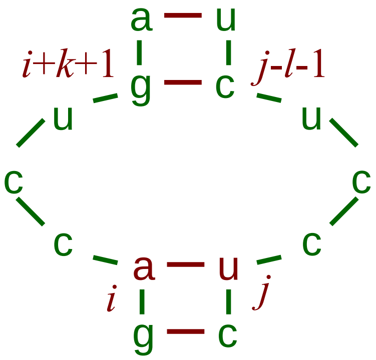
$$P(i,j)=\min\begin{cases}\alpha(i,j)+\text{LoopEnergy}(j-i-1)\\ \alpha(i,j)+\text{StackingEnergy}(i,j,i+1,j-1)+P(i+1,j-1)\\ \min\limits_{k\geq 1}\left[\alpha(i,j)+\text{BulgeEnergy}(k)+P(i+k+1,j-1)\right]\\ \min\limits_{k\geq 1}\left[\alpha(i,j)+\text{BulgeEnergy}(k)+P(i+1,j-k-1)\right]\\ \min\limits_{k,l\geq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

# Predicting RNA Secondary Structure by Energy Minimization

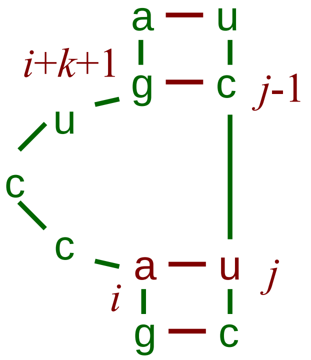
$$\alpha(i,j)+\text{LoopEnergy}(j-i-1)$$



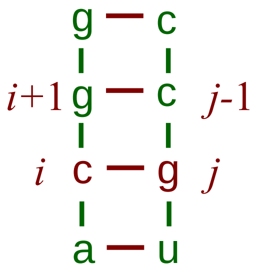
$$\min_{k,l\geq 1}\left[\alpha(i,j)+\text{LoopEnergy}(k+l)+P(i+k+1,j-l-1)\right]$$



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

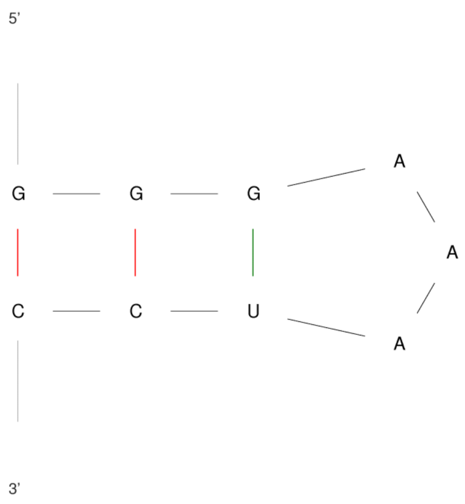


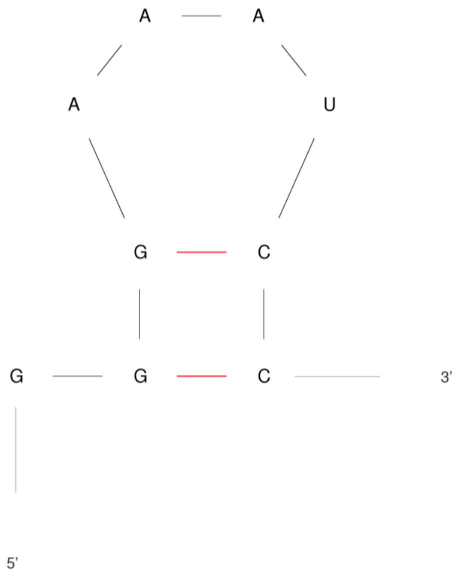
$$\alpha(i,j)+\text{StackingEnergy}(i,j,i+1,j-1)+P(i+1,j-1)$$



# Mfold example

**GGGAAAUCC**





ΔG = -0.80 kcal/mol

ΔG = 0.20 kcal/mol

http://unafold.rna.albany.edu/

# Summary

- RNA has numerous roles in
  - translation, splicing, DNA replication, gene regulation
- RNA structure understanding is important
  - substitutions are possible, function preserved as long as they preserve the structure
- Secondary structure can be predicted
  - comparative sequence analysis
    - molecules with similar function will form similar structures
    - searches for positions that co-vary
  - free energy minimization
    - take single sequence, search for energetićally stable complementary regions
    - in a simplified form discussed in this lecture
    - current folding programs get about 50-70% base pairs correct on average
    - a large number of foldings lie close to the predicted global energy minimum
  - in general an intractable task