

RNA Secondary Structure Prediction

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Mark Craven

craven@biostat.wisc.edu

Spring 2009

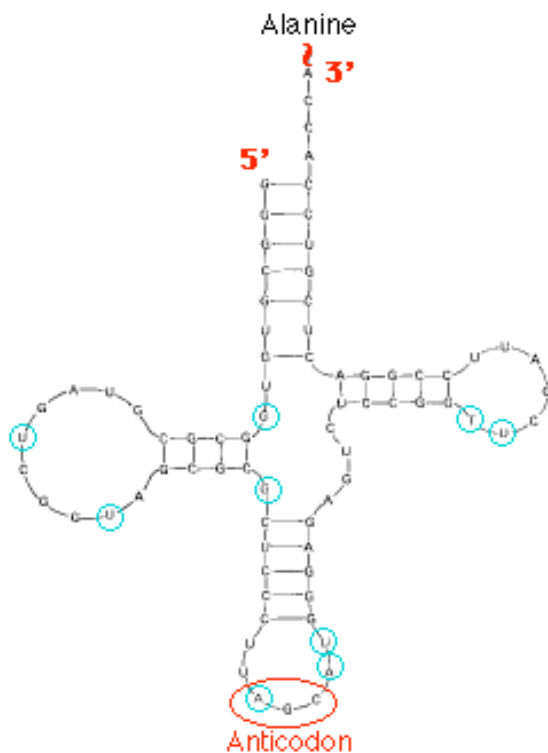
Why RNA Is Interesting

- in addition to messenger RNA (mRNA), there are other RNA molecules that play key roles in biology
 - 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
 - the genomes for many viruses (e.g. HIV) are encoded in RNA
 - etc.
- the folding of an mRNA can be involved in regulating the gene's expression

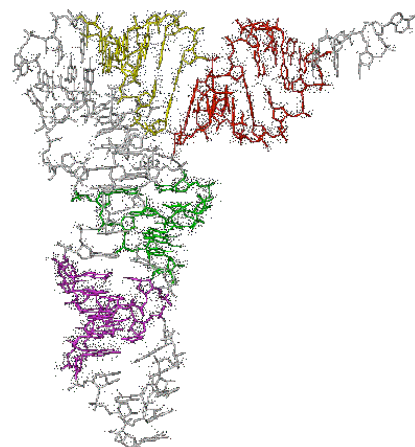
RNA Secondary Structure

- 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**
- the 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



tertiary structure



Domain: *Bacteria*
Kingdom: *Purple Bacteria*
Order: *gamma*

Four Key Problems

- predicting RNA secondary structure
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

- the 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

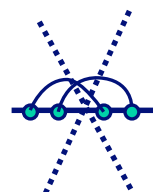


$i < i' < j < j'$ or $j < j' < i < i'$ (not nested)



$i < j < j' < i'$ 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

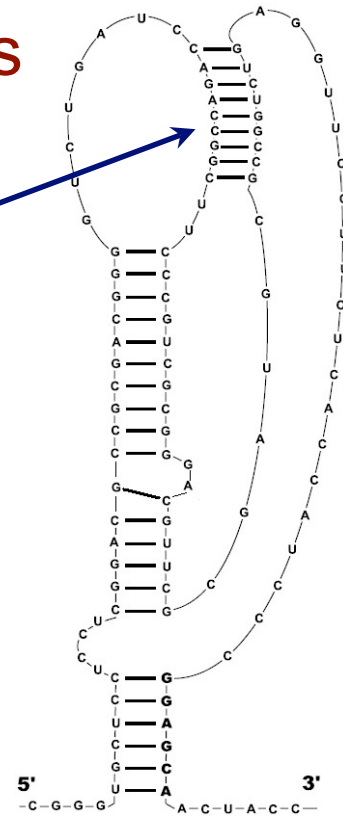


Figure from Seliverstov et al. *BMC Microbiology*, 2005

Predicting RNA Secondary Structure: the Nussinov Algorithm

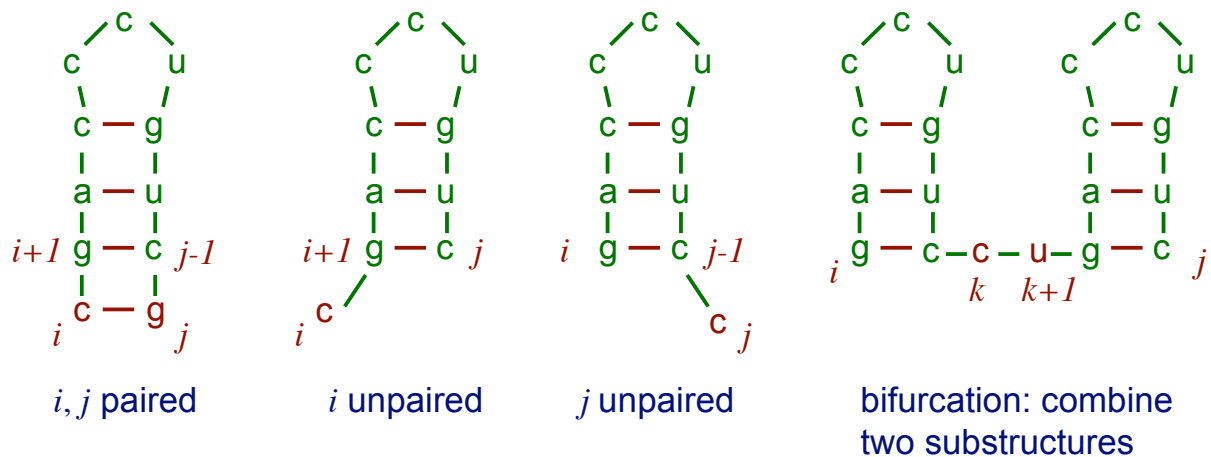
Nussinov et al., *SIAM Journal of Applied Mathematics* '78

key idea:

- find maximal number of base pairings for a given sequence
- do this using dynamic programming
 - start with small subsequences
 - progressively work to larger ones

DP in the Nussinov Algorithm

- incrementally find the best secondary structure for subsequences $[i, j]$
- consider four possibilities



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:

$$\begin{aligned} \gamma(i, i-1) &= 0 & \text{for } i = 2 \text{ to } L \\ \gamma(i, i) &= 0 & \text{for } i = 1 \text{ to } L \end{aligned}$$

- 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

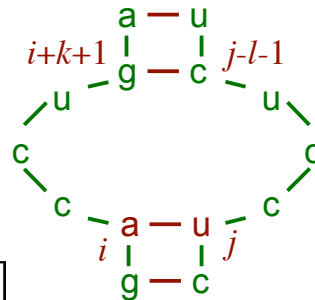
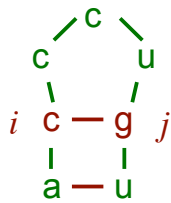
Predicting RNA Secondary Structure by Energy Minimization

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

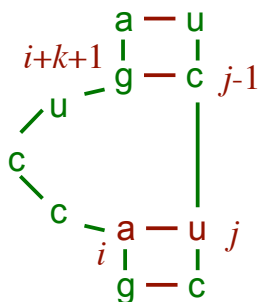
$$P(i, j) = \min \begin{cases} \alpha(i, j) + \text{LoopEnergy}(j - i - 1) \\ \alpha(i, j) + \text{StackingEnergy}(i, j, i + 1, j - 1) + E(i + 1, j - 1) \\ \min_{k \geq 1} [\alpha(i, j) + \text{BulgeEnergy}(k) + E(i + k + 1, j - 1)] \\ \min_{k \geq 1} [\alpha(i, j) + \text{BulgeEnergy}(k) + E(i + 1,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

Predicting RNA Secondary Structure by Energy Minimization

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



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



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

