RNA secondary structure prediction

Jiří Kléma

Department of Computer Science,
Czech Technical University in Prague

Lecture based on Mark Craven's class at University of Wisconsin

IDA
Intelligent Data Analysis
RESEARCH GROUP

http://cw.felk.cvut.cz/wiki/courses/b4m36bin/start

Overview

- Key concepts
  - RNA secondary structure,
  - secondary structure features: stems, loops, bulges,
  - Nussinov algorithm,
  - adapting Nussinov to take free energy into account.
- untouched
  - special base pairs: non-canonical, base triplets, pseudoknots,
  - advanced algorithms including deep networks, transfer learning etc.

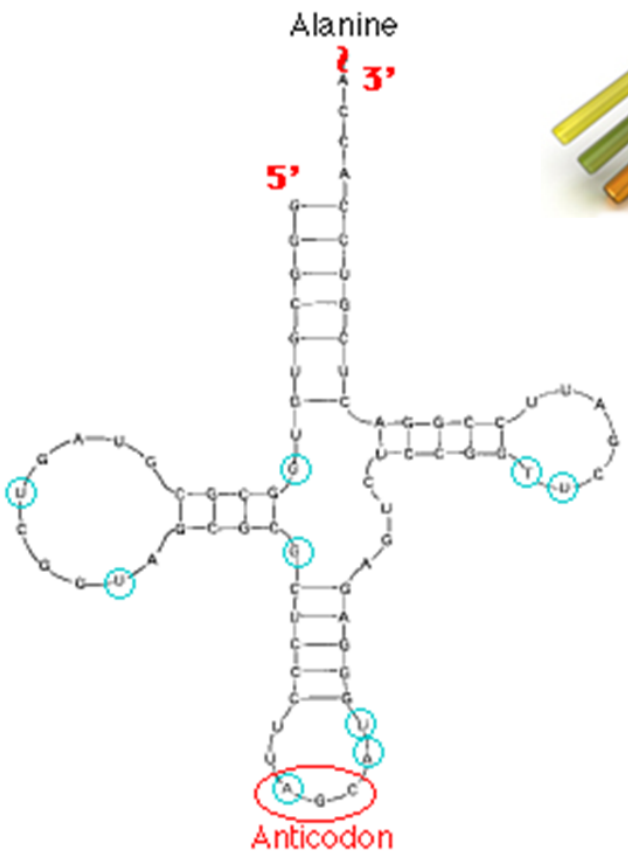
Why RNA is interesting

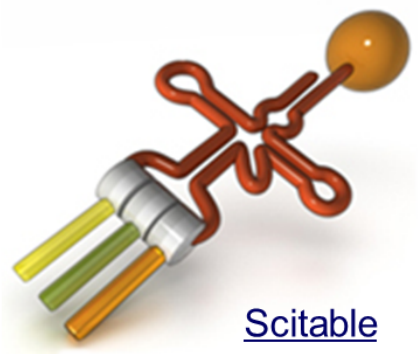
- Messenger RNA (mRNA) is not the only important class of RNA
  - ribosomal RNA (rRNA)
    - ribosomes are complexes that incorporate several RNA subunits in addition to numerous protein units,
  - transfer RNA (tRNA)
    - transport amino acids to the ribosome during translation,
  - the spliceosome,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

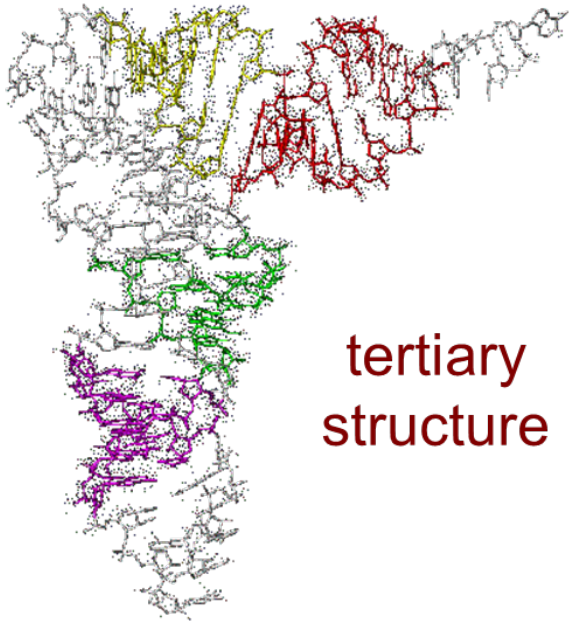
RNA secondary structure

- 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

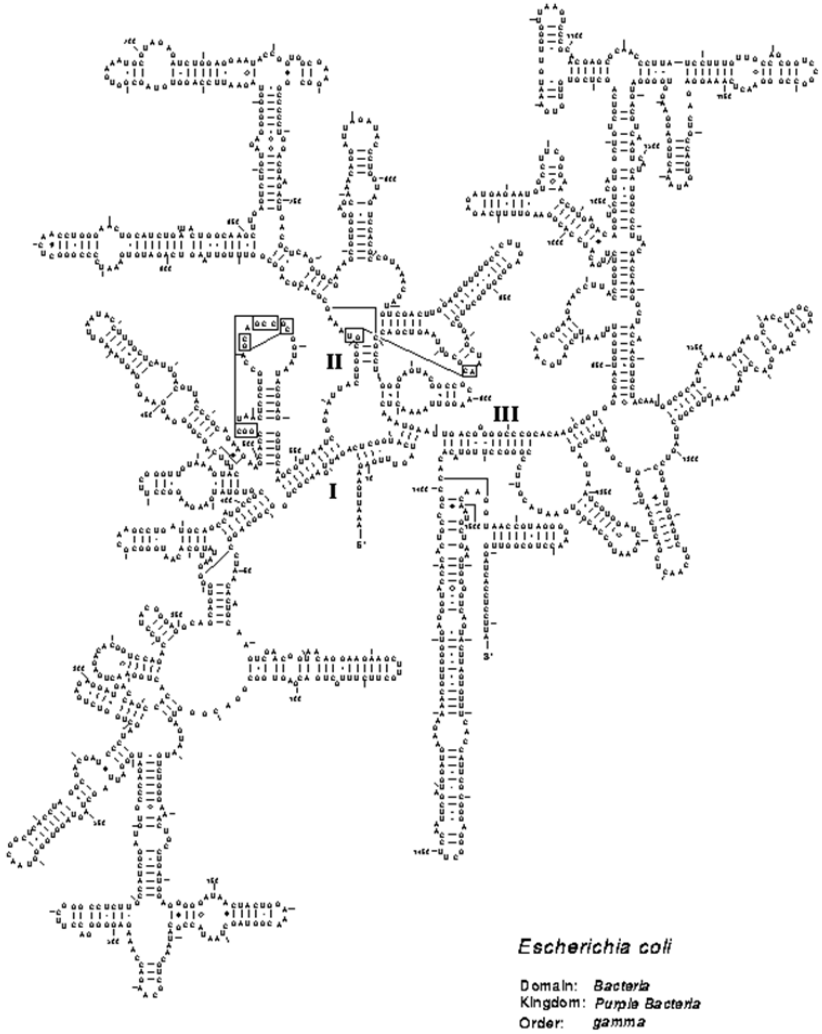






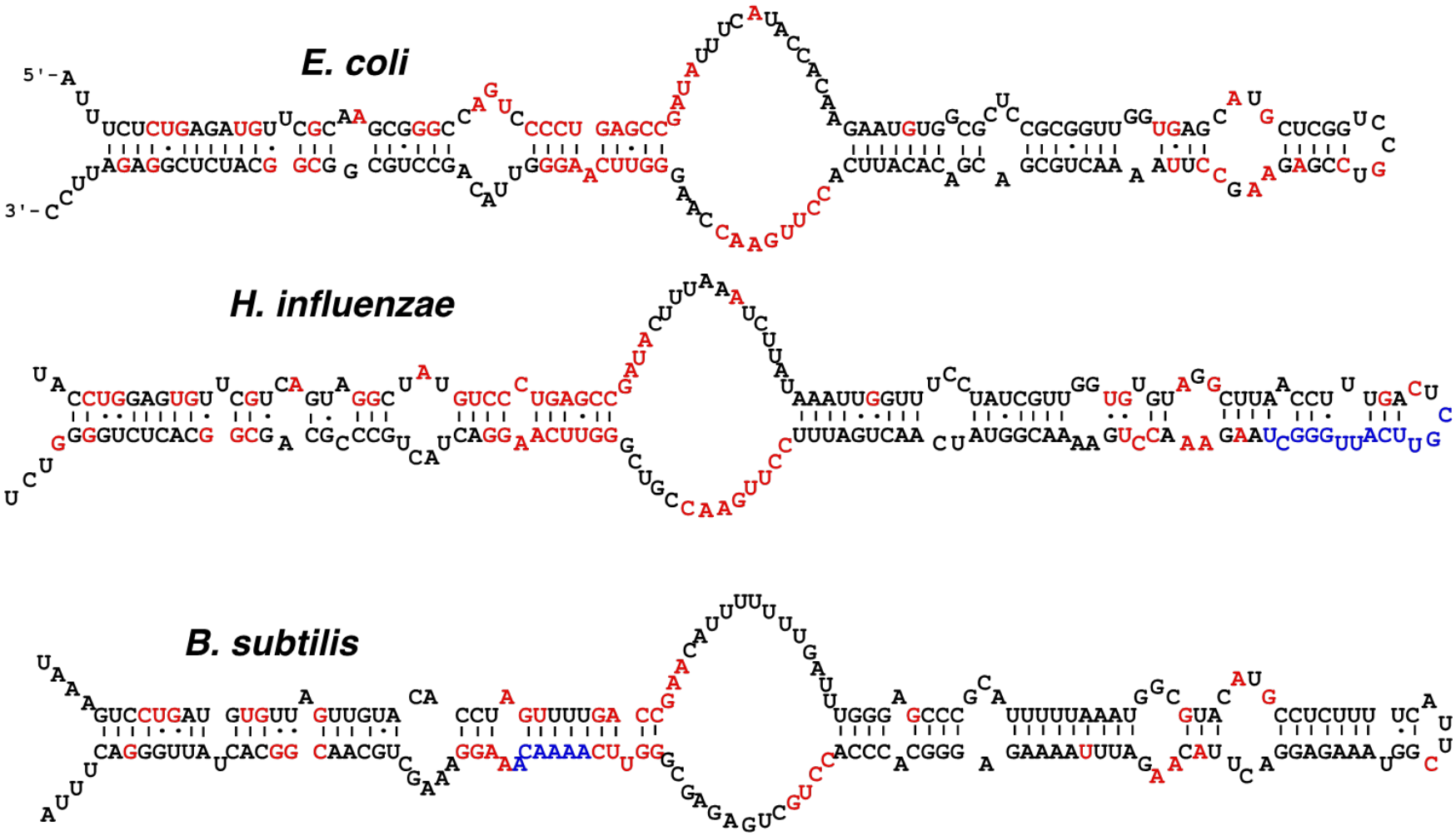
Marc Craven, BMI/CS 576, www.biostat.wisc.edu/bmi576.

Small subunit ribosomal RNA



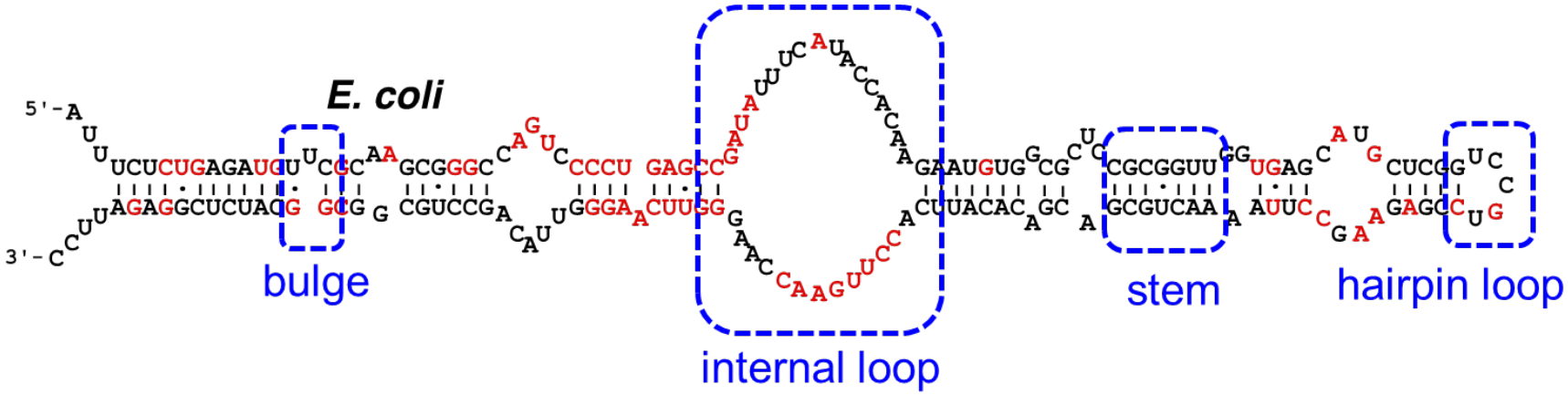
O'Connor, Nucleic acids research, 1997.

6S RNA secondary structure



Cavanagh, Annual review of microbiology, 2014.

Secondary structure features

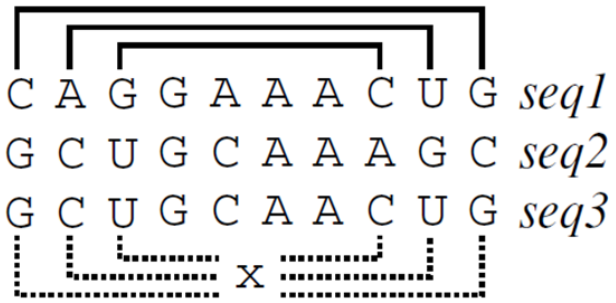


Marc Craven, BMI/CS 576, www.biostat.wisc.edu/bmi576.

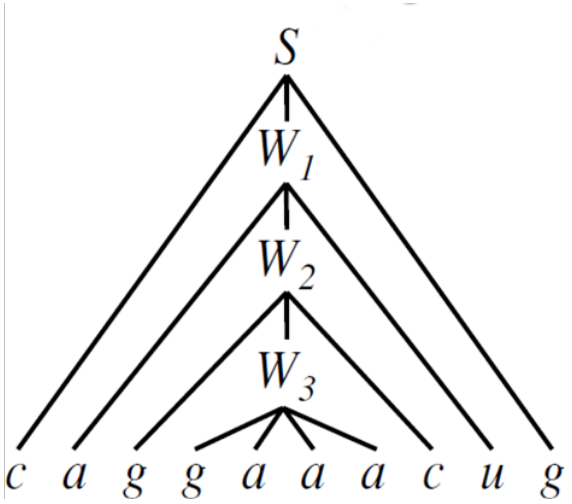
Secondary structure as CFG

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

seq1	*seq2*	*seq3*
A A	C A	C A
G A	G A	G A
G • C	U • A	U × C
A • U	C • G	C × U
C • G	G • C	G × G



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



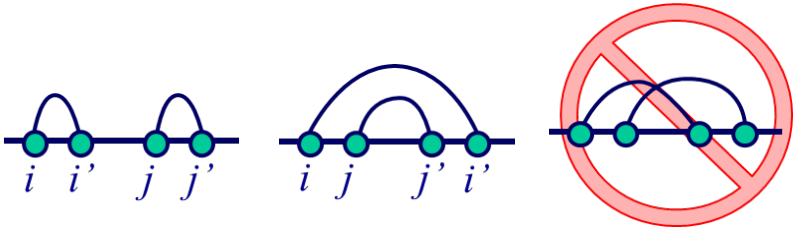
Durbin, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.

Four key problems

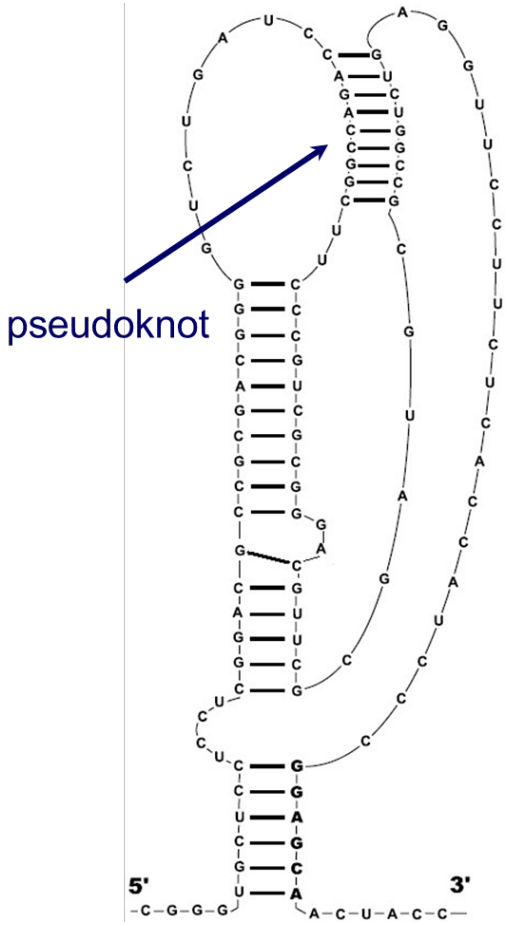
- Predicting RNA secondary structure (**Focus for today**)
  - Given: RNA sequence,
  - Do: predict secondary structure that sequence will fold into,
- Searching for instances of a given structure
  - Given: an RNA sequence or its secondary structure,
  - Do: find sequences that will fold into a similar 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 and pseudoknots

- We will 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),
  - or $i < j < j' < i'$ or $j < i < i' < j'$ (nested),
- cannot have $i < j < i' < j'$ or $j < i < j' < i'$
  - these crossings are called **pseudoknots**,
  - dynamic programming breaks down with them,
  - fortunately, they are not very frequent.



not nested nested crossing



Seliverstov et al. BMC Microbiology, 2005.

Predicting RNA secondary structure

- Given:
  - an RNA sequence,
  - the constraint = pseudoknots not allowed,
- Do:
  - find a secondary structure for the RNA,
  - it maximizes the number of base pairing positions,
- Nussinov algorithm
  - key ideas
    - do this using dynamic programming,
    - start with small subsequences,
    - progressively work to larger ones.

DP in the Nussinov algorithm

- Let

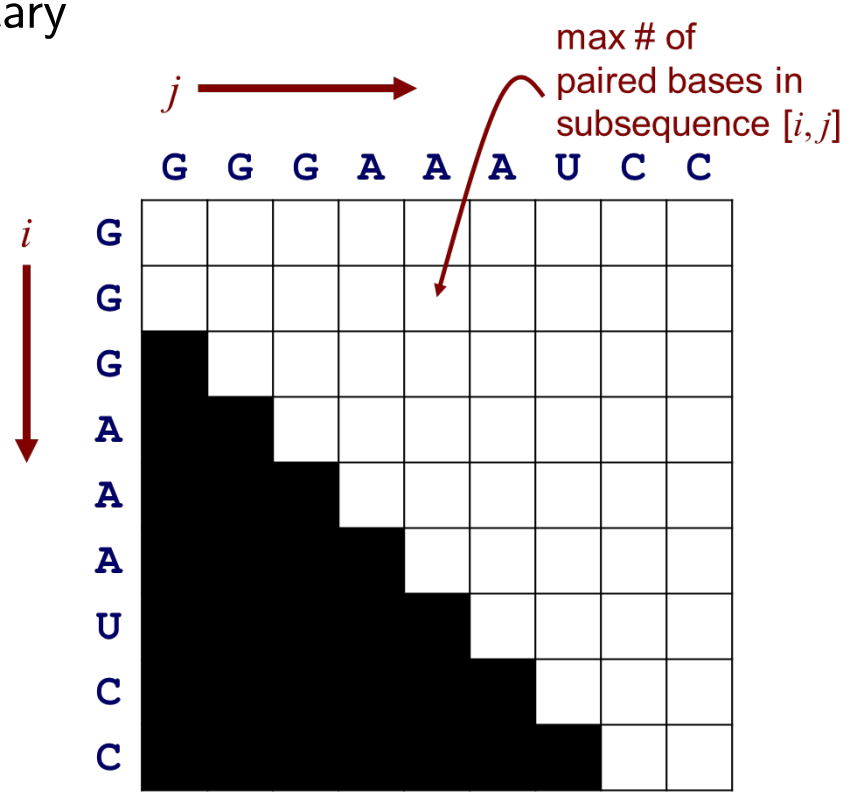
$$\delta(i,j) = \begin{cases} 1 \text{ if } x_i \text{ and } x_j \text{ complementary} \\ 0 \text{ otherwise} \end{cases}$$

- 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

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



Durbin, Biological Sequence Analysis:
Probabilistic Models of Proteins and Nucleic Acids.

Nussinov algorithm traceback

- Determine one non-crossing RNA structure with maximal score.

```
push(1, L) onto stack
repeat until stack is empty
  pop(i, j)
  if i ≥ j continue
  else if γ(i + 1, j) = γ(i, j) push(i + 1, j)
  else if γ(i, j − 1) = γ(i, j) push(i, j − 1)
  else if γ(i + 1, j − 1) + δ(i, j) = γ(i, j)
    record i, j base pair
    push(i + 1, j − 1)
  else for k = i + 1 to j − 1:
    if γ(i, k) + γ(k + 1, j) = γ(i, j)
      push(k + 1, j)
      push(i, k)
      break
```

Predict RNA secondary structure by energy minimization

- Maximizing the number of base pairs oversimplifies prediction of folding,
- however, we can generalize the key recurrence relation by minimizing free energy instead.

$$E(i,j) = min \begin{cases} E(i+1,j) \\ E(i,j-1) \\ min_{i<k<j}[E(i,k) + E(k+1,j)] \\ P(i,j) \quad \leftarrow \quad \text{case that } i \text{ and } j \text{ are base paired} \end{cases}$$

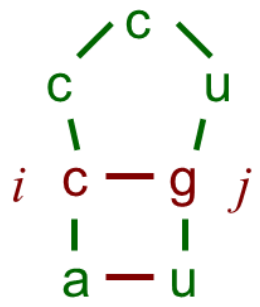
Predict RNA secondary structure by energy minimization

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

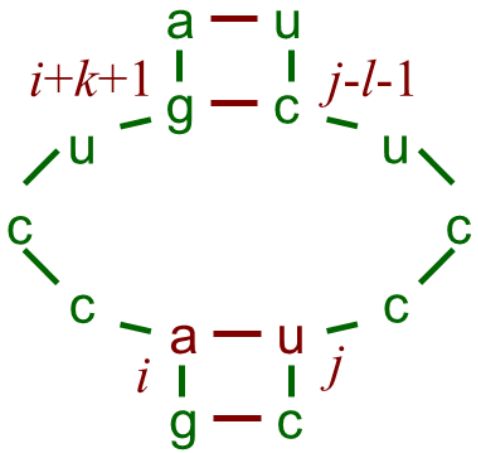
$$P(i,j) = min \begin{cases} \alpha(i,j) + \text{LoopEnergy}(j-i-1) \\ \alpha(i,j) + \text{StackingEnergy}(i,j,i+1,j-1) + P(i+1,j-1) \\ min_{k\geq 1}[\alpha(i,j) + \text{BulgeEnergy}(k) + P(i+k+1,j-1)] \\ min_{k\geq 1}[\alpha(i,j) + \text{BulgeEnergy}(k) + P(i+1,j-k-1)] \\ min_{k,l\geq 1}[\alpha(i,j)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

Predict RNA secondary structure by energy minimization

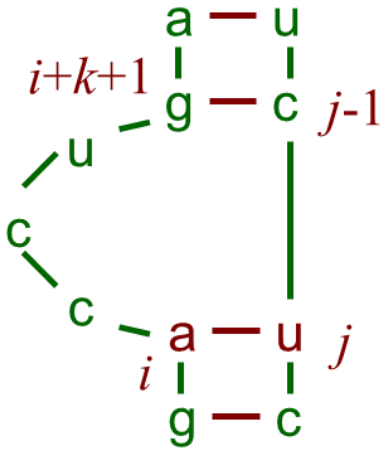
$\alpha(i,j) + \text{LoopEnergy}(j-i-1)$



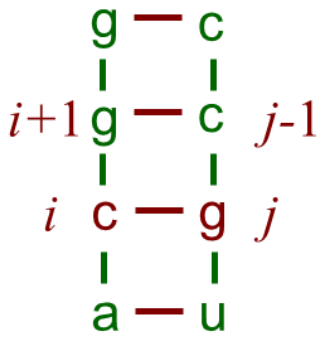
$min_{k,l \geq 1}[\alpha(i,j) + \text{LoopEnergy}(k+l) + P(i+k+1, j-l-1)]$



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



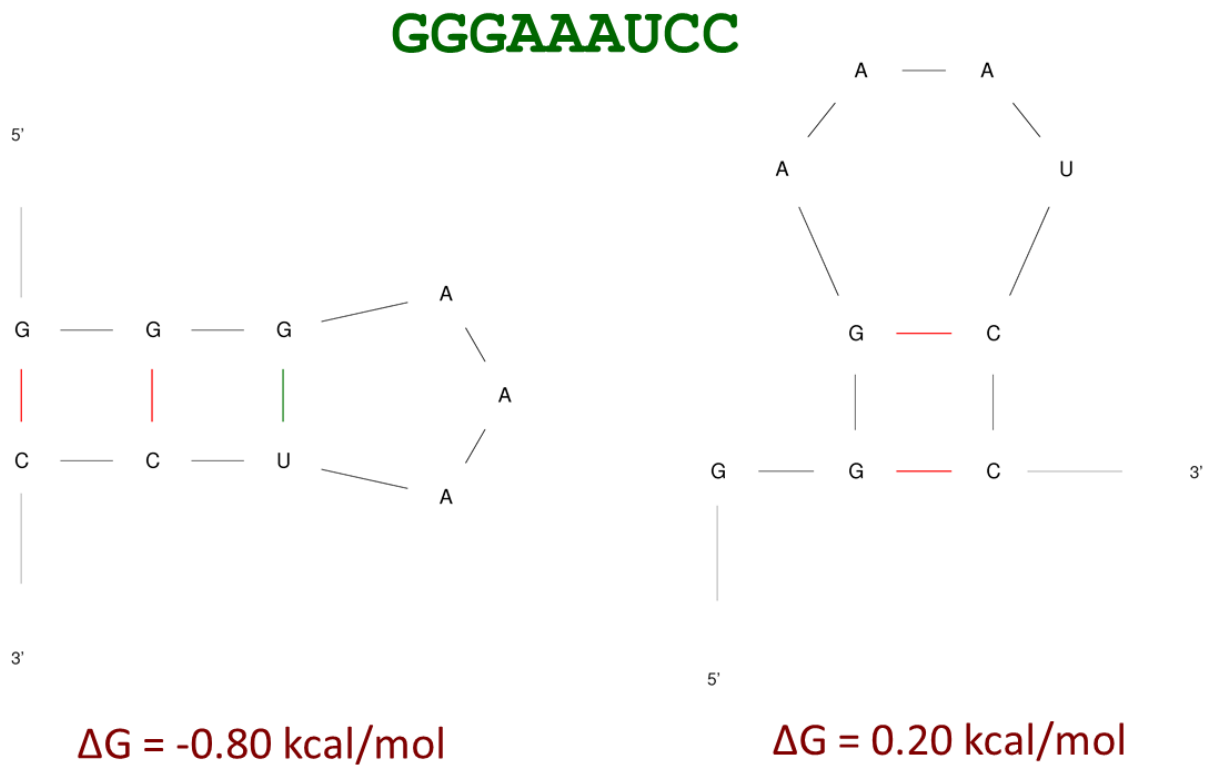
$\alpha(i,j) + \text{StackingEnergy}(i,j,i+1,j-1) + P(i+1,j-1)$



Marc Craven, BMI/CS 576, www.biostat.wisc.edu/bmi576.

Mfold example

- Mfold solutions with energy up to 5% from the best
  - different from Nussinov results (2 Watson-Crick base pairs only here).



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

Summary

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