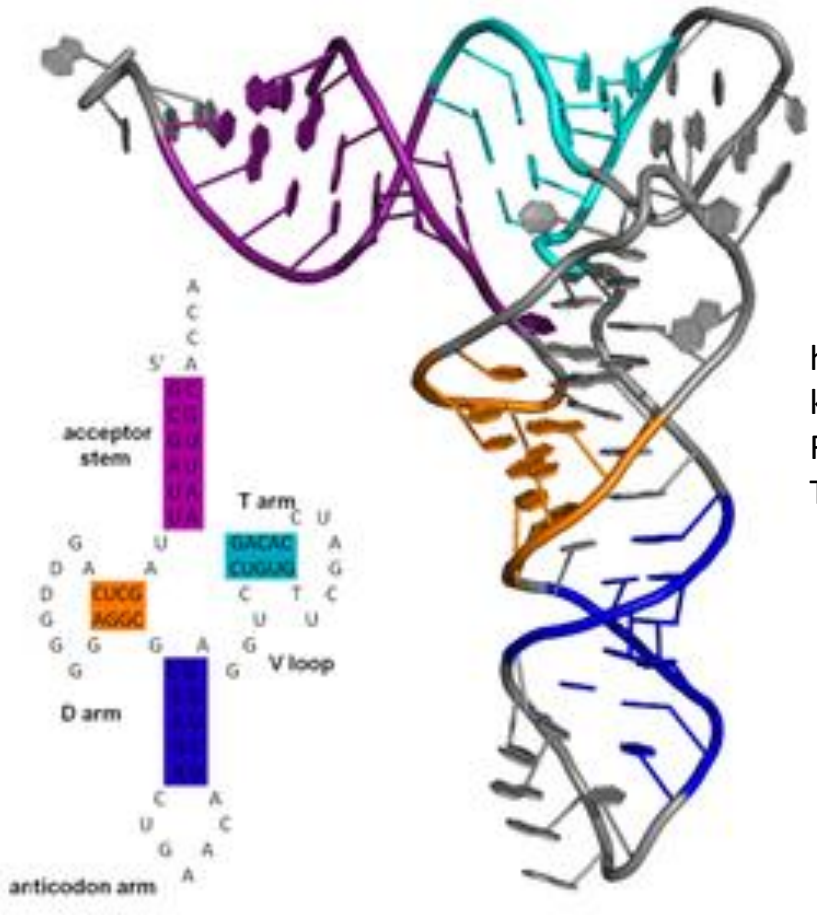
EECS730: Introduction to Bioinformatics

Lecture 09: Non-coding RNA structure prediction



https://upload.wikimedia.org/wikipedia/commons/thumb/b/bf/TRNA_all2.png/250px-TRNA_all2.png

Slides adapted from Dr. Shaojie Zhang (University of Central Florida)

microRNA

Tiny RNAs–"Biological Equivalent of Dark Matter"–Wins Prestigious AAAS Newcomb Cleveland Prize

The discovery of micro-sized RNA molecules (miRNAs)–a breakthrough described as "the biological equivalent of dark matter, all around us but almost escaping detection"–earned the coveted 2001-2002 AAAS Newcomb Cleveland Prize.

Three journal reports, published in the 26 October 2001 issue of *Science*, were named to receive the Prize, the oldest award conferred by the American

C. elegans

- miRNAs were the second major story in 2001 (after the genome).
- Subsequently, many other non-coding genes have been found

The Nobel Prize in Physiology or Medicine 2006

"for their discovery of RNA interference - gene silencing by double-stranded RNA"

Photo: L. Cicero/Stanford

Photo: R. Carlin/UMMAS

Andrew Z. Fire

1/2 of the prize

USA

Stanford University School of Medicine
Stanford, CA, USA

b. 1959

Craig C. Mello

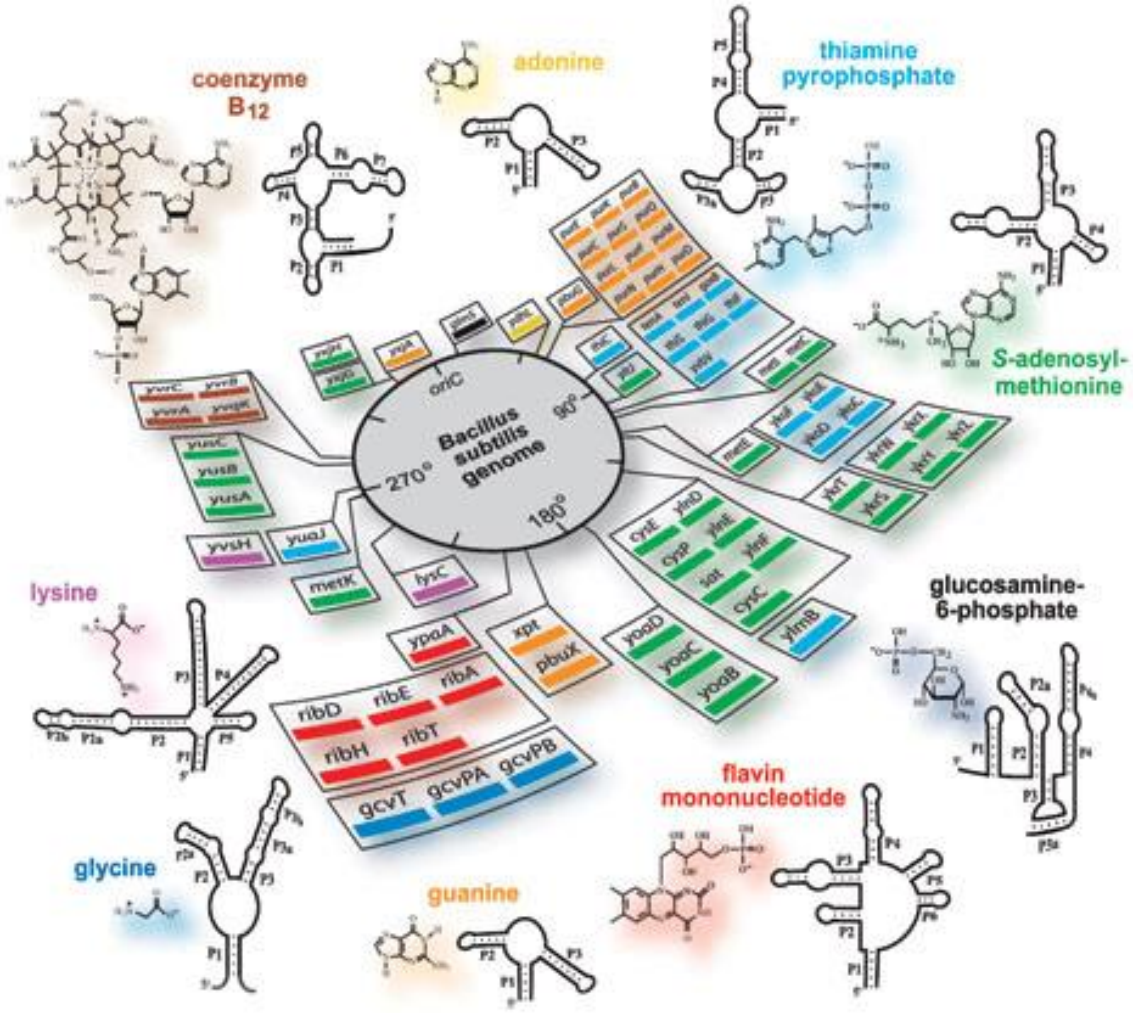
1/2 of the prize

USA

University of Massachusetts Medical School
Worcester, MA, USA

b. 1960

riboswitches



- -Breaker Lab

Decoding the genome

- Current gene prediction methods only work well for protein coding genes.
- Non-coding RNA genes are undetected because they do not encode proteins.
- Modern RNA world hypothesis:
  - There are many unknown but functional ncRNAs. [Eddy *Nature Reviews* (2001)]
  - Many ncRNAs may play important role in the unexplained phenomenon.[Storz *Science* (2002)]

Noncoding RNAs

RNA

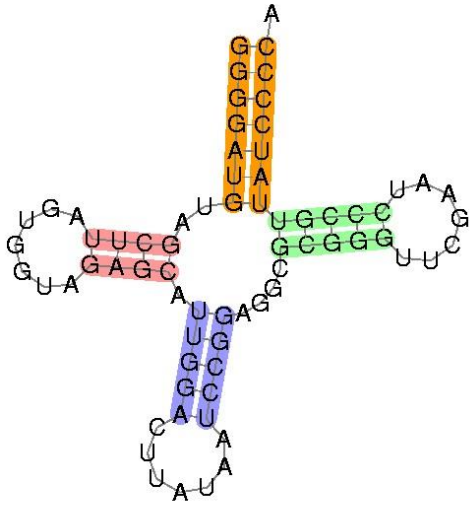
We sequenced RNA[16] from different cell lines and multiple subcellular fractions to develop an extensive RNA expression catalogue. Using a conservative threshold to identify regions of RNA activity, 62% of genomic bases are reproducibly represented in sequenced long (>200 nucleotides) RNA molecules or GENCODE exons. Of these bases, only 5.5% are explained by GENCODE exons. Most transcribed bases are within or overlapping annotated gene boundaries (that is, intronic), and only 31% of bases in sequenced transcripts were intergenic[16].

The ENCODE Project Consortium, *Nature*, 2007

Noncoding RNAs

- Non-coding RNA (ncRNA)
  - RNA acting as functional molecule.
  - Not translated into protein.
- The RNA world hypothesis:
  - RNA are as important as protein coding genes.
  - Many undiscovered ncRNA exist
- Computational methods for discovering ncRNA are not mature.
- What are the clues to non-coding genes?
  - Structure: Given a sequence, what is the structure into which it can fold with minimum energy?

RNAs conserve on structure rather than sequence



Compensatory mutation

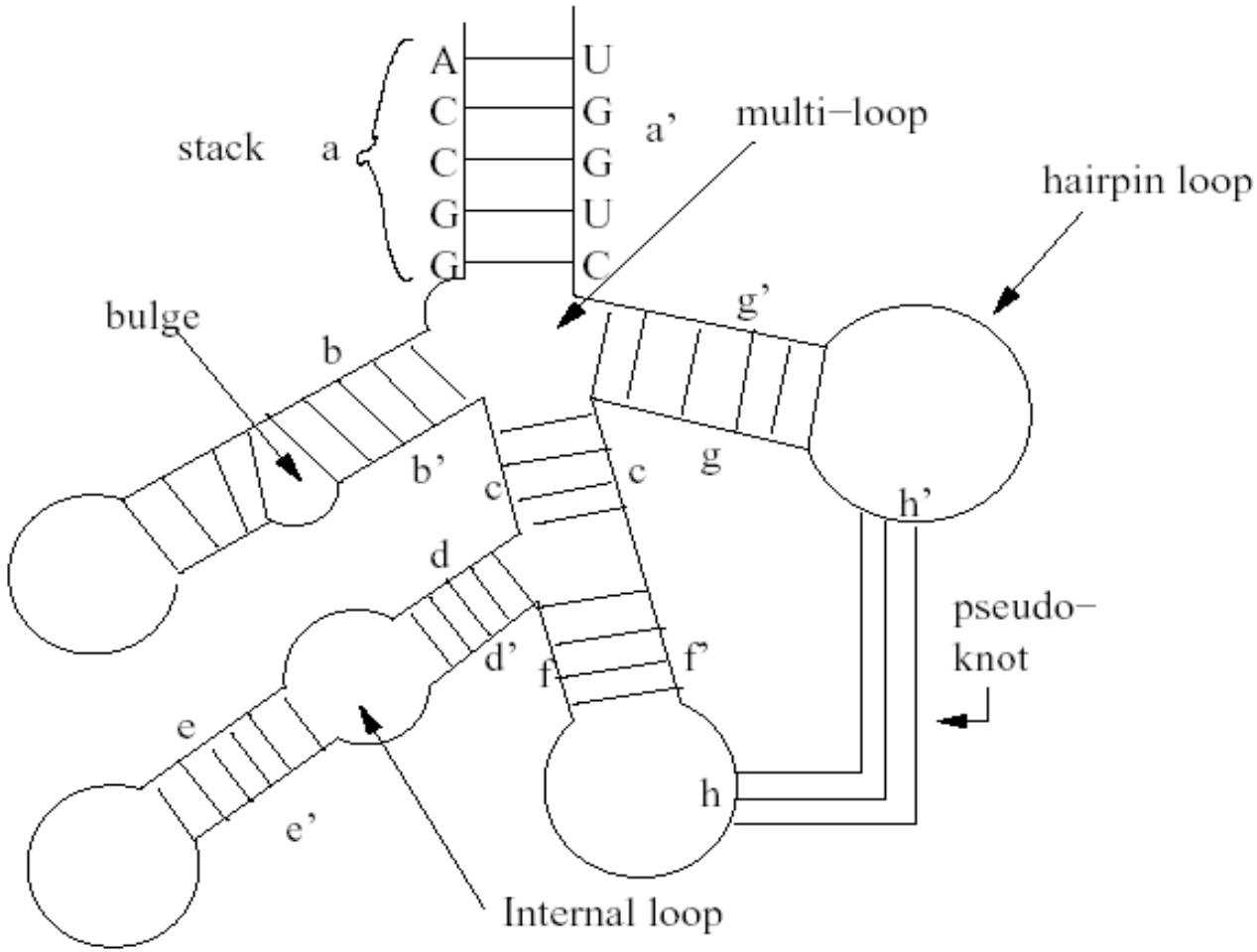
```
tRNA1: GCAUCGGUGGUUCAGUGGUAGAAUGCUCGCCUGCCACGCGGGCG
       <<<<<<<..<<<<.......>>>>.<<<<<.......>>>>>..
tRNA2: UCUAAUAUGGCAGAUU...AGUGCAAUAGAUUUAAGCUCUAUAU

       GCCCGGGUUCGAUUCCCGACCGAUGCA
       ..<<<<<.......>>>>>>>>>>>.
       AUAAAGU.AUUUU.ACUUUUAUUAGAA
```

http://www.sanger.ac.uk/Software/Rfam/

RNA structure basis

- Key: RNA is single-stranded. Think of a string over 4 letters, A, C, G, and U.
- The complementary bases form pairs.
  - A <-> U, C <-> G, G <-> U
- Base-pairing defines a secondary structure. The base-pairing is usually non-crossing.
- Functional biomolecules are often energetically stable, i.e. they have low minimum free energy (MFE)

Components of RNA structure



The RNA secondary structure folding problem

- ncRNA is not a random sequence.
- Most RNAs fold into particular base-paired secondary structure.
- Finding the set of base pairs that minimize the free energy
- Most basepairs are non-crossing basepairs.
- Any two pairs (i, j) and (i',j'): $i < i' < j' < j$ or $i' < i < j < j'$ or $i < j < i' < j$ or $i' < j' < i < j$

- Canonical basepairs:
  - Watson-Crick basepairs:
    - G - C
    - A - U
  - Wobble basepair:
    - G – U

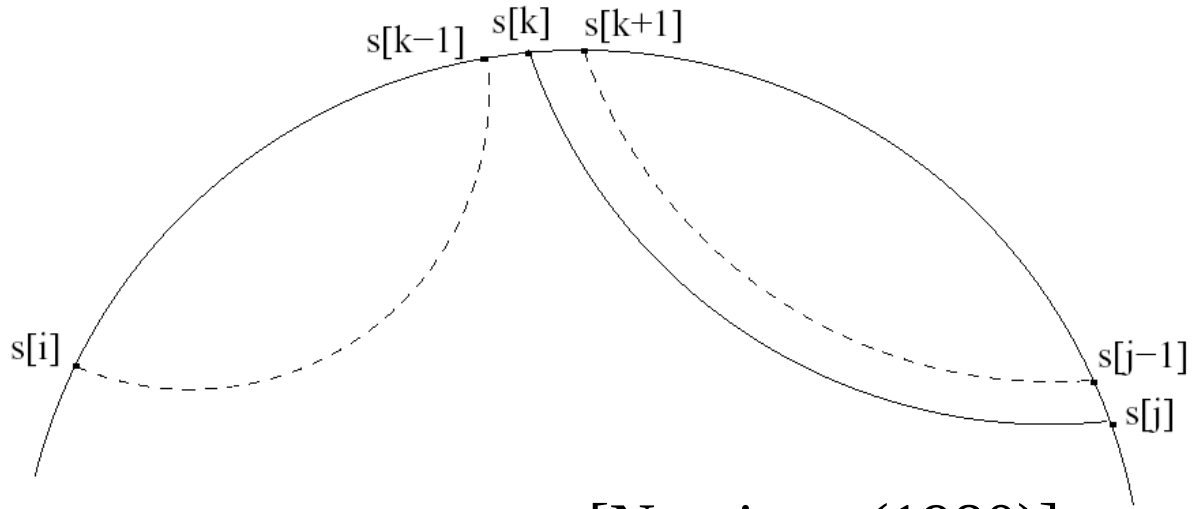
Naïve formulation

- A simple energy model is to maximize the number of basepairs to minimize the free energy. [Waterman (1978), Nussinov et al (1978), Waterman and Smith (1978)]
- G – C, A – U, and G – U are treated as equal stability.
- Contributions of stacking are ignored.

Problem 1: [Base pair maximization problem]
Given an RNA sequence, determine a set of base pairs in a RNA sequence such that the number of base pairs is maximal and no base pairs cross each other.

Dynamic programming solution

- Let $s[1…n]$ be an RNA sequence.
- $\delta(i,j) = 1$ if $s[i]$ and $s[j]$ form a complementary base pair, else $\delta(i,j) = 0$.
- $M(i,j)$ is the maximum number of base pairs in $s[i…j]$.



[Nussinov (1980)]

Dynamic programming solution

$$M(i,j) = max \begin{cases} M(i,j-1), \\ M(i,k-1) + M(k+1,j-1) + \delta(k,j) \\ \qquad\qquad\qquad\qquad \text{for } i \leq k < j. \end{cases}$$

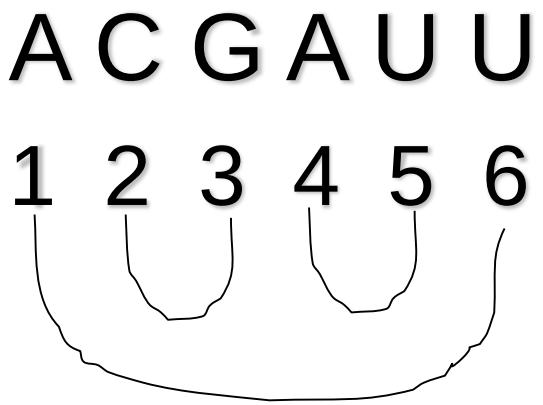
- *M*(*1,n*) is the number of base pairs in the optimal base-paired structure for *s*[*1…n*].
- All these base pairs can be found by tracing back through the matrix *M*.
- Time complexity of the algorithm?

Time complexity of the algorithm

- O(n^3), for all i, j, and k

An example

j \ i	1	2	3	4	5	6
2	0					
3	1	1				
4	1	1	0			
5	2	2	1	1		
6	3	2	1	1	0	

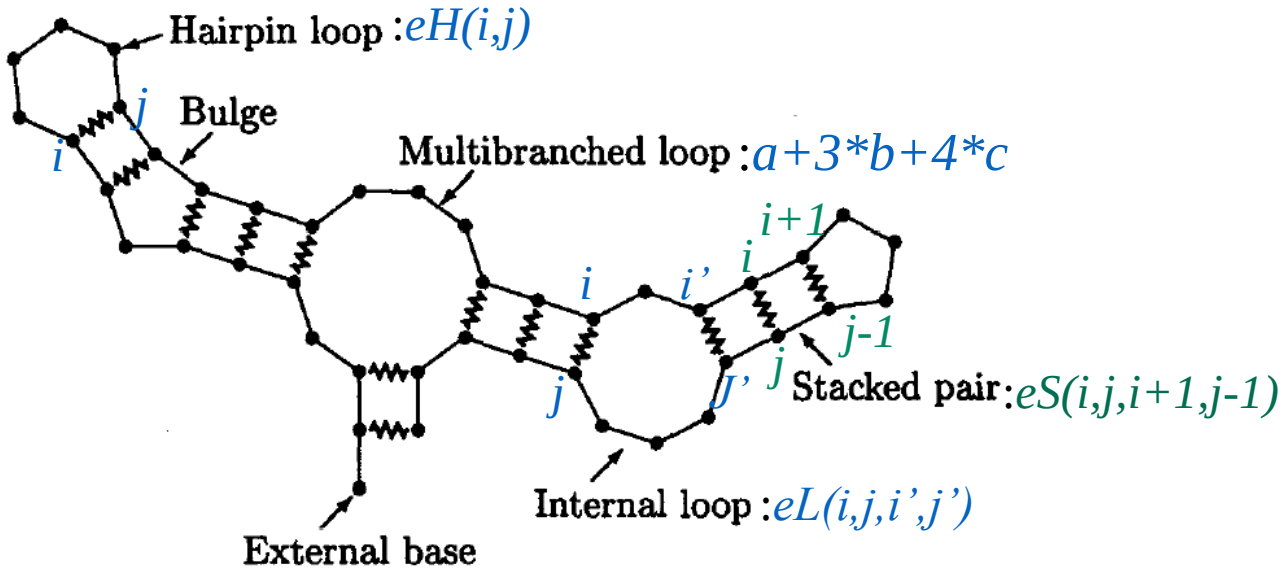


Zuker-Sankoff Model

- Stacks (contiguous nested base pairs) are the dominant stabilizing force – contribute the negative energy
- Unpaired bases form loops contribute the positive energy.
  - Hairpin loops, bulge/internal loops, and multiloops.

- Zuker-Sankoff minimum energy model. [Zuker and Sankoff (1984), Sankoff (1985)]
- `Mfold` and `ViennaRNA` are all based on this model. (this model is also called mfold model)

Zuker-Sankoff MFE model



[Lyngsø (1999)]

- **Hairpin loop:** $eH(i,j)$ is the energy of the hairpin loop from $i+1$ to $j-1$, which *closed* by base pair (i,j).
- **Stacked base pairs:** $eS(i,j,i+1,j-1)$ is the energy of the stacking base pairs (i,j) and $(i+1,j-1)$.
- **Bulge and internal loop:** $eL(i,j,i',j')$ is the energy of the the bulge or internal loop starting from $i+1$ to $i'-1$ and from $j'+1$ to $j'-1$ which is *closed* by base pairs (i,j) and (i',j').
- **Multi-loop:** a is the energy of generating a multi-loop, b is the energy of one base pair that *closes* the multi-loop, and c is the energy of one unpaired base in the multi-loop.

Time complexity?

- O(n^4), why?
- We can reduce it to O(n^3) by slightly modifying the energy function

Recursive function (Zuker-Sankoff)

- *W*(*i*) holds the minimum energy of a structure on *s*[*1…i*].
- *V*(*i,j*) holds the minimum energy of a structure on *s*[*i…j*] with *s*[*i*] and *s*[*j*] forming a basepair.
- *WM*(*i,j*) holds the minimum energy of a structure on *s*[*i…j*] that is part of multiloop.

$$W(i) = \min\{W(i-1), \min_{0 \le k < i}\{W(k) + V(k+1, i)\}\}.$$

$$V(i,j) = \min\{eH(i,j), eS(i,j,i+1,j-1) + V(i+1,j-1), \min_{i<i'<j'<j \text{ and } i'-i+j-j'>2}\{eL(i,j,i',j') + V(i',j')\}, \min_{i+1<k<j}\{WM(i+1,k-1) + WM(k,j-1) + a\}\},$$

$$WM(i,j) = \min\{V(i,j) + b, WM(i,j-1) + c, WM(i+1,j) + c, \min_{i<k\le j}\{WM(i,k-1) + WM(k,j)\}\},$$

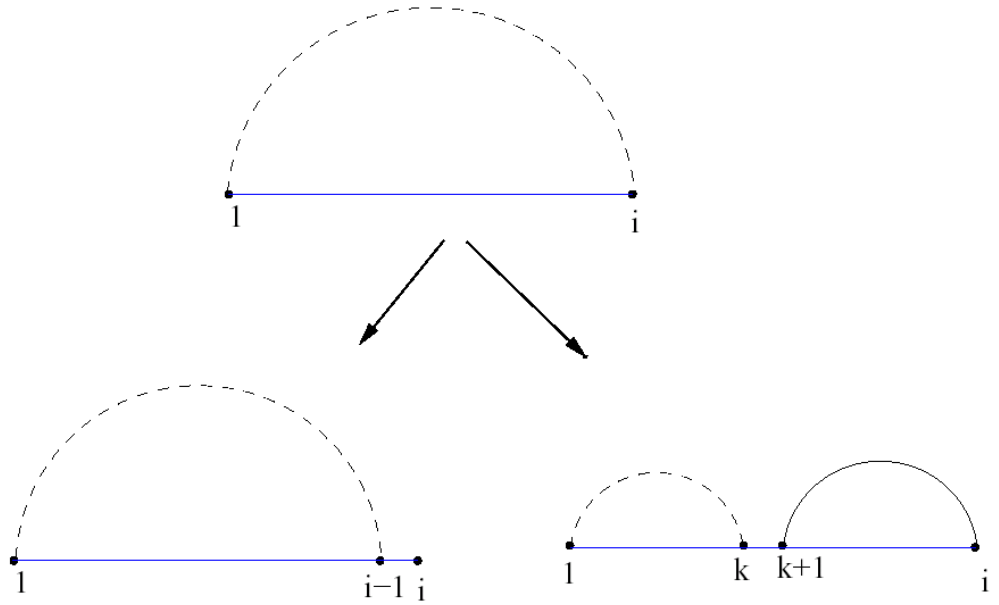
Recursive function (Zuker-Sankoff)

- *W*(*i*) holds the minimum energy of a structure on *s*[*1…i*].

$$W(i) = \min\{W(i-1), \min_{0 \le k < i}\{W(k) + V(k+1, i)\}\}.$$

- *V*(*i,j*) holds the minimum energy of a structure on *s*[*i…j*] with *s*[*i*] and *s*[*j*] forming a basepair.

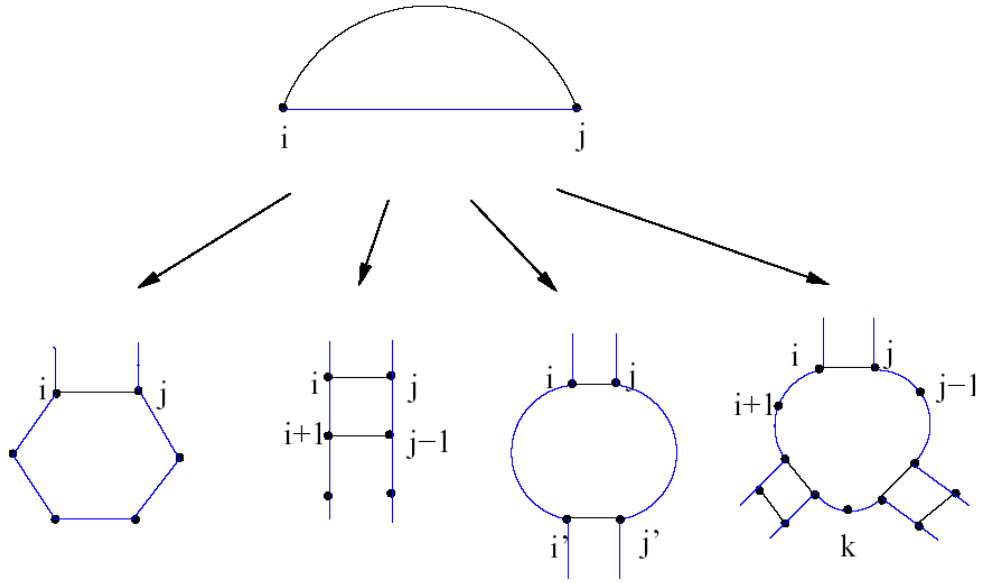
- *WM*(*i,j*) holds the minimum energy of a structure on *s*[*i…j*] that is part of multiloop.



Recursive function (Zuker-Sankoff)

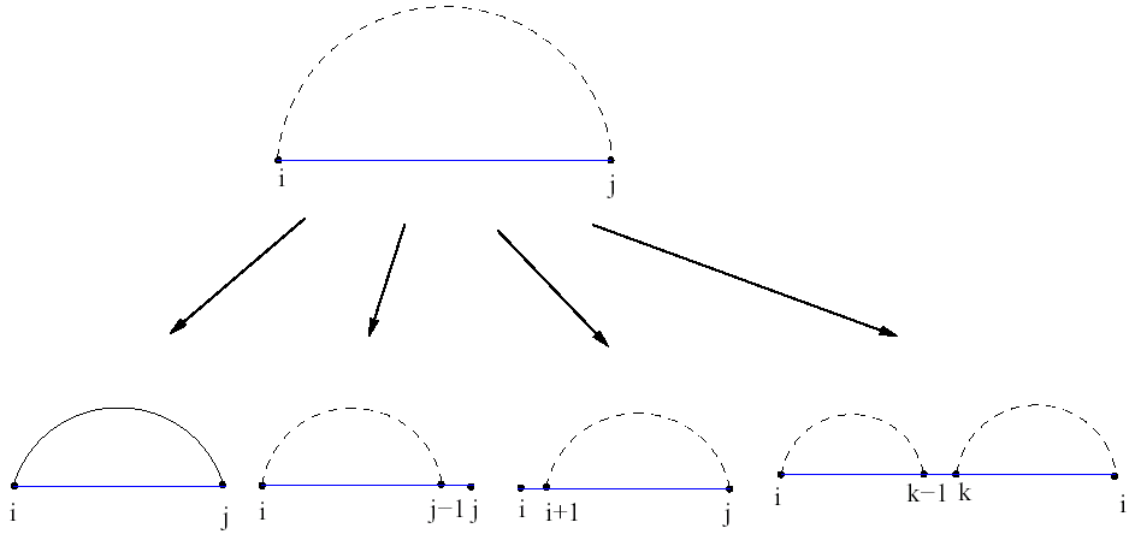
- *W*(*i*) holds the minimum energy of a structure on *s*[*1…i*].
- *V*(*i,j*) holds the minimum energy of a structure on *s*[*i…j*] with *s*[*i*] and *s*[*j*] forming a basepair.
- *WM*(*i,j*) holds the minimum energy of a structure on *s*[*i…j*] that is part of multiloop.

$$V(i,j) = \min\{eH(i,j),$$
$$eS(i,j,i+1,j-1) + V(i+1,j-1),$$
$$\min_{i<i'<j'<j \text{ and } i'-i+j-j'>2} \{eL(i,j,i',j') + V(i',j')\},$$
$$\min_{i+1<k<j} \{WM(i+1,k-1) + WM(k,j-1) + a\}\},$$



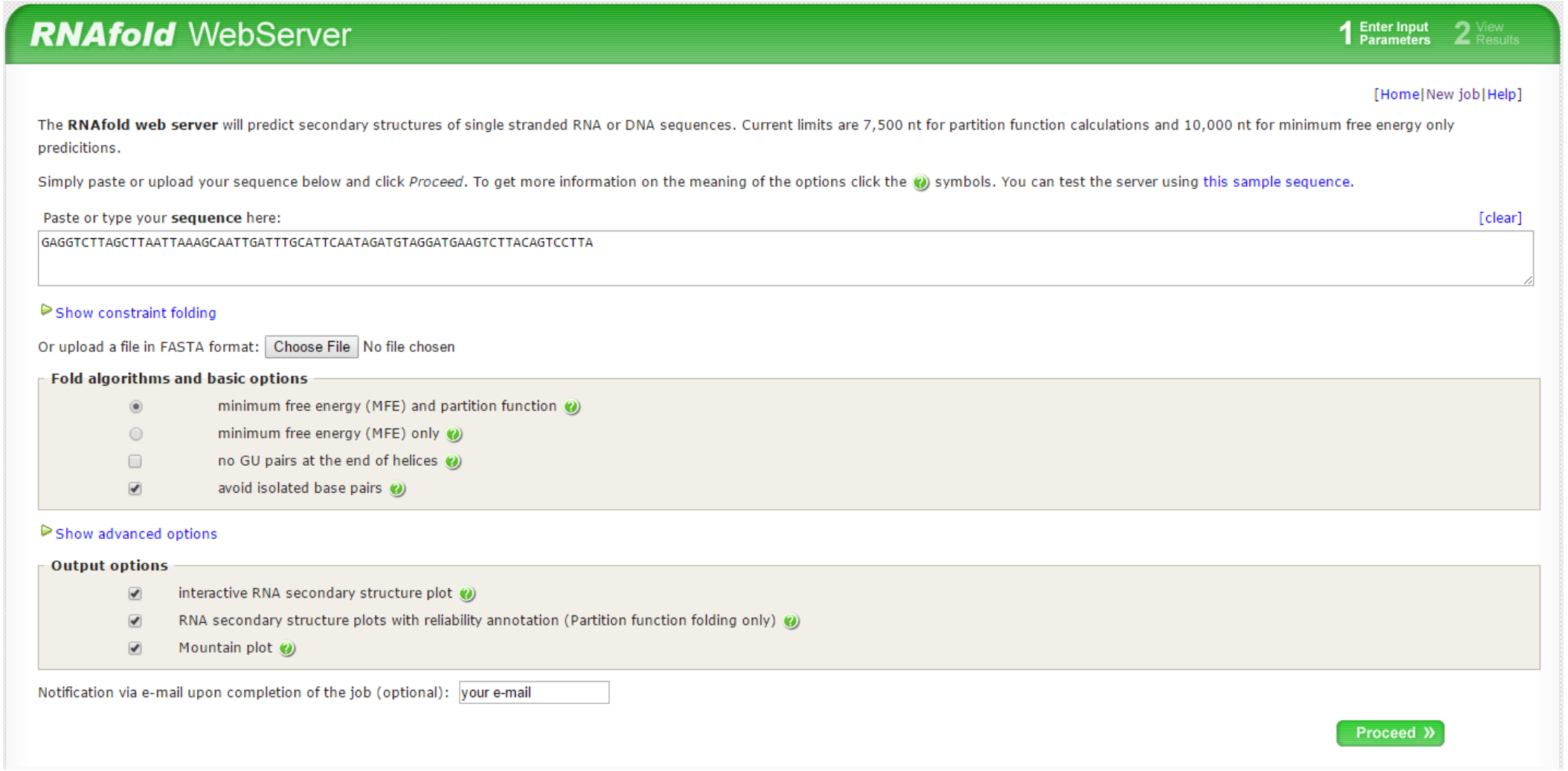
Recursive function (Zuker-Sankoff)

- *W*(*i*) holds the minimum energy of a structure on *s*[*1…i*].
- *V*(*i,j*) holds the minimum energy of a structure on *s*[*i…j*] with *s*[*i*] and *s*[*j*] forming a basepair.
- *WM*(*i,j*) holds the minimum energy of a structure on *s*[*i…j*] that is part of multiloop.



$$
\begin{aligned}
WM(i,j) = \min\{ & V(i,j) + b, \\
& WM(i,j-1) + c, \\
& WM(i+1,j) + c, \\
& \min_{i<k\leq j} \{WM(i,k-1) + WM(k,j)\}\},
\end{aligned}
$$

RNAfold server



GAGGTCTTAGCTTAATTAAAGCAATTGATTTGCATTCAATAGATGTAGGATGAAGTCTTACAGTCCTTA