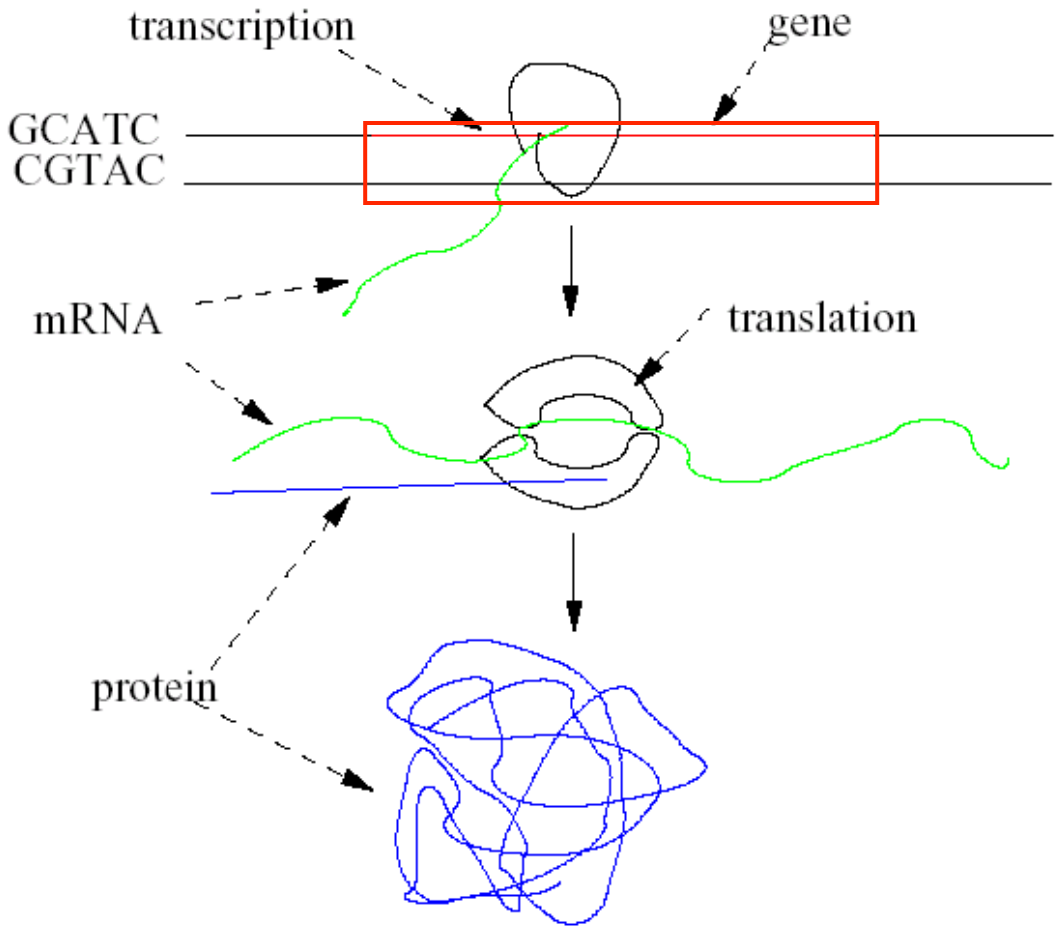
Take-home Problem: the number of alignments

- Let A(n,m) = the number of alignments of one sequence with length i and one sequence with length j.
- A(n,m)=A(n-1,m)+A(n-1,m-1)+A(n,m-1).
- $$A(m,n) = \sum_{k=0}^{\min(m,n)} \frac{(m+n-k)!}{k!(m-k)!(n-k)!}$$

		n										
		0	1	2	3	4	5	6	7	8	9	10
m	0	1										
	1	1	3									
	2	1	5	13								
	3	1	7	25	63							
	4	1	9	41	129	321						
	5	1	11	61	231	681	1683					
	6	1	13	85	377	1289	3653	8989				
	7	1	15	113	575	2241	7183	19825	48639			
	8	1	17	145	833	3649	13073	40081	108545	265729		
	9	1	19	181	1159	5641	22363	75517	224143	598417	1462563	
	10	1	21	221	1561	8361	36365	134245	433905	1256465	3317445	8097453

Non-coding RNA gene finding problems

Central dogma cont'd

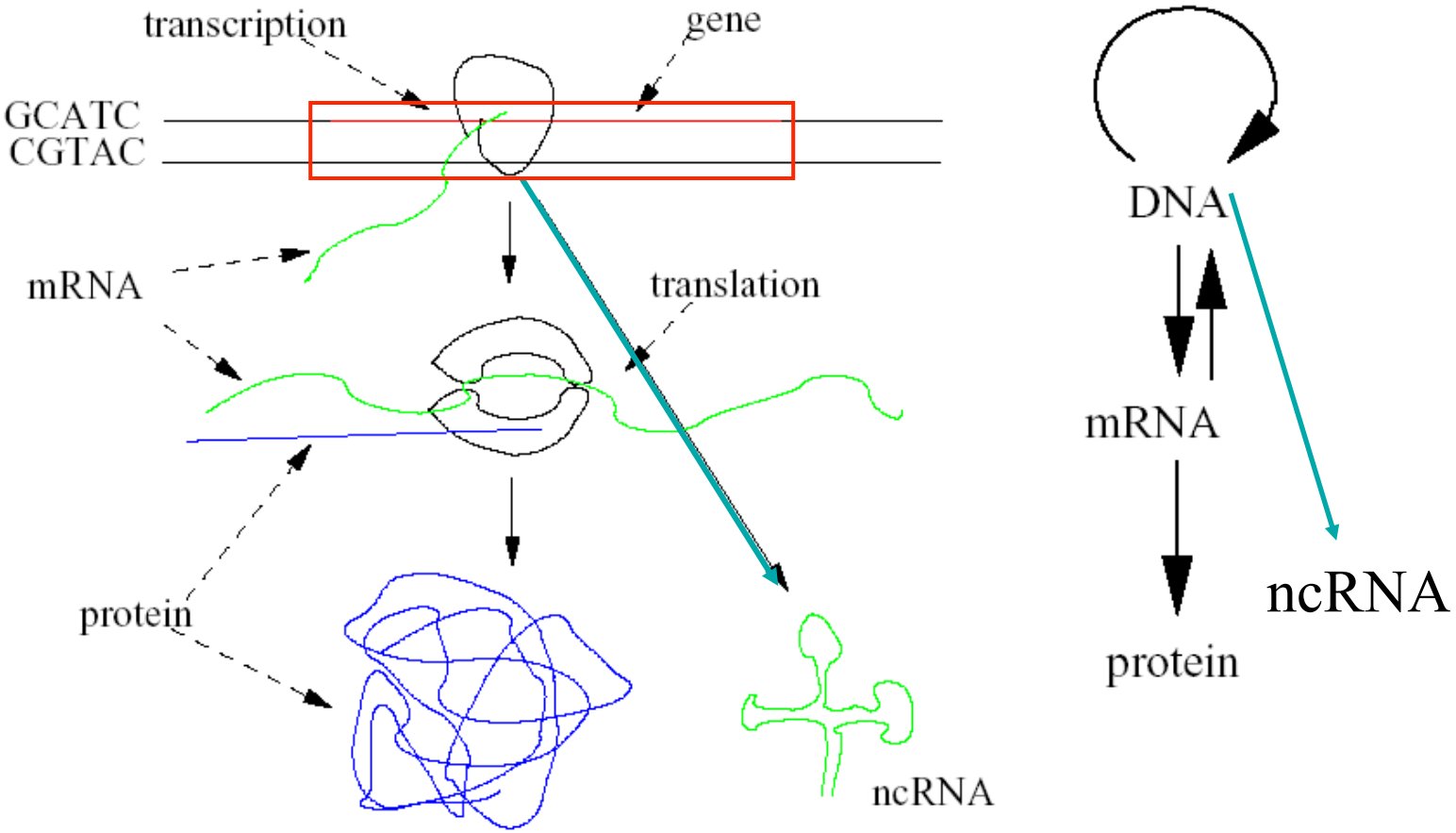


Central dogma cont'd

- Non-coding RNA (ncRNA)
  - RNA acting as functional molecule.
  - Not translated into protein.

- Non-coding RNA gene
  - The region of DNA coding ncRNA.

Central dogma cont'd



Human genome

How many genes do we have?

- Only about 30,000 to 40,000 protein-coding genes in the human genome [Lander et al. *Nature* (2001), Venter et al. *Science* (2001)] .

- Total protein coding gene length is only about 1.5 percent of the human genome. ($3*10^9$ bases)

What did we miss out?

- 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)]

Question:

If there are many ncRNAs, what are they doing?

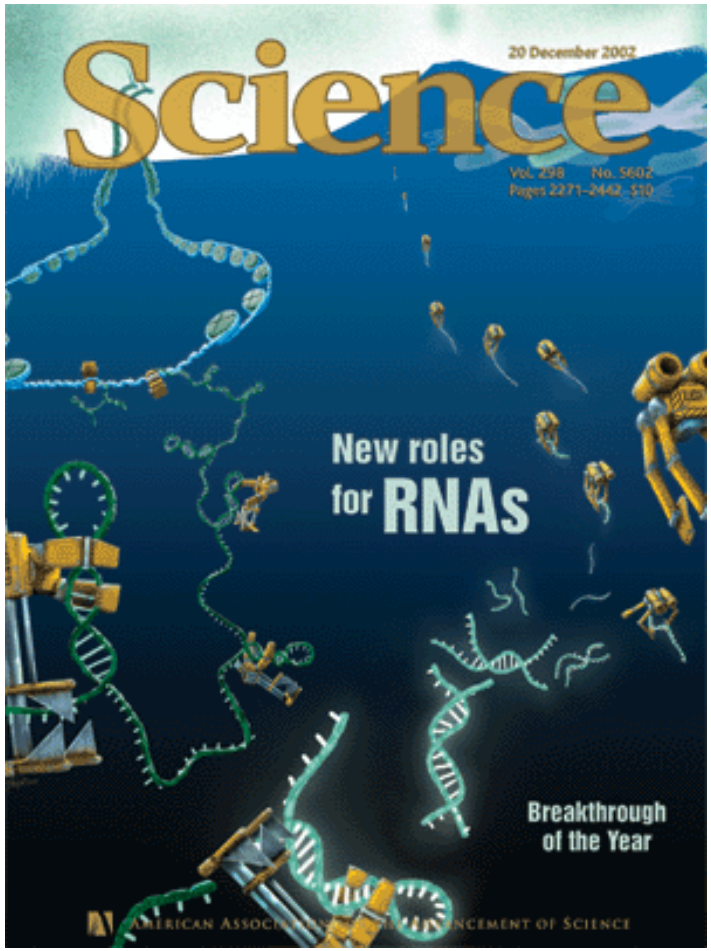
Question:

Biologically, why do we need functional ncRNAs in addition to protein?

Why do we need ncRNAs?

- ncRNAs involve sequence specific recognition of other nucleic acids (e.g. mRNAs, DNAs).
- ncRNA is an ideal material for this role.
  - DNA is big and packaged and can do this job.
- *Base complementary* allows ncRNA to be sequence specific!
- For example:
  - small interfering RNAs (siRNA) is used to protect our genome.
  - It recognizes invading foreign RNAs/DNAs based on the sequence specificity.
  - And helps to degrade the foreign RNAs.


Science
17 May 2002
Vol. 296 - No. 5571
Pages 1189–1352 $9
RNA Silencing and
Noncoding RNA
American Association for the Advancement of Science


Science
20 December 2002
Vol. 298 No. 5602
Pages 2271–2442 $10
New roles
for RNAs
Breakthrough
of the Year

How can we find such ncRNA genes in the genome?

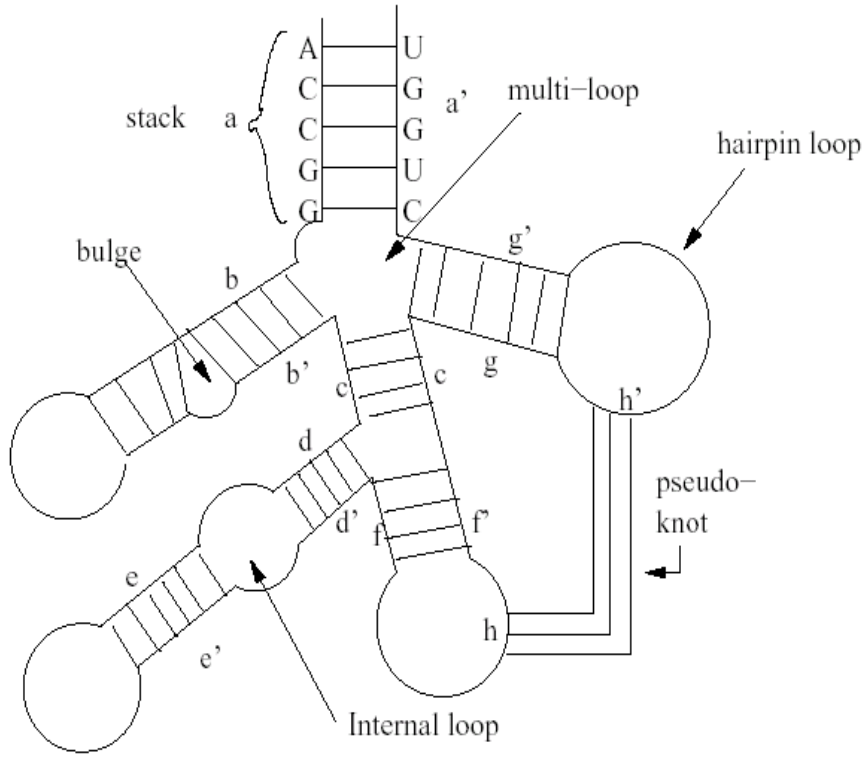
RNA secondary structure

- ncRNA is not a random sequence.
- Most RNAs fold into particular base-paired secondary structure.

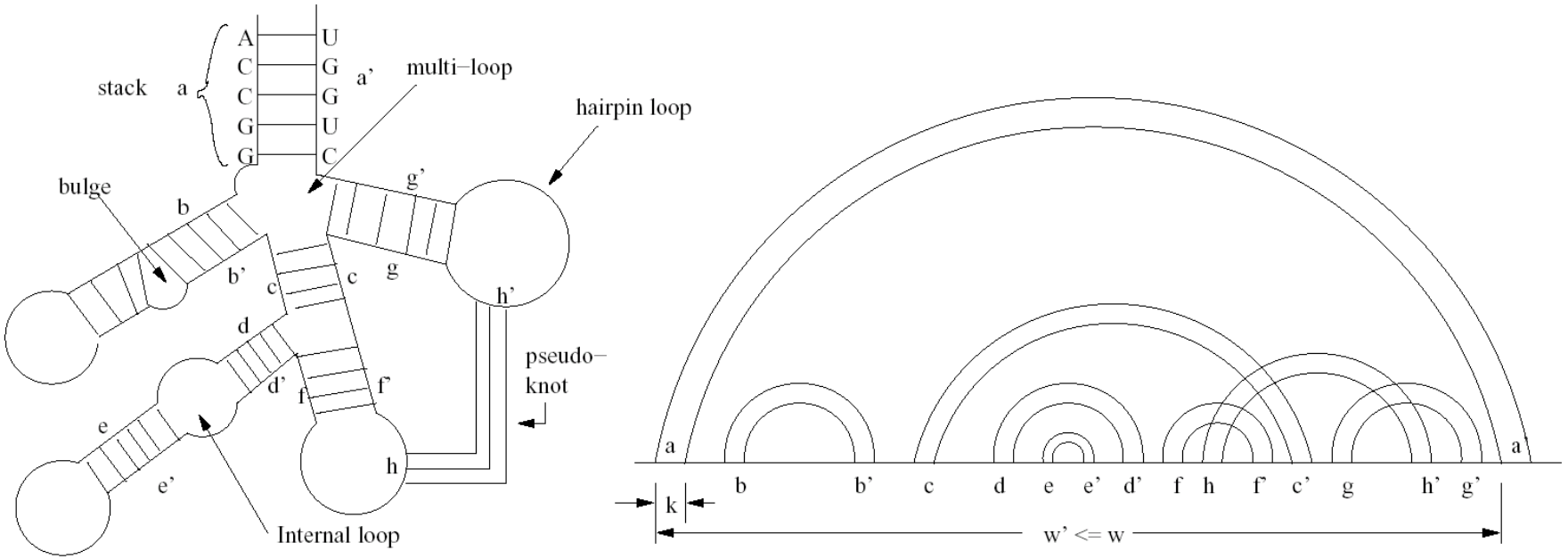
- Canonical basepairs:
  - Watson-Crick basepairs:
    - G - C
    - A - U
  - Wobble basepair:
    - G – U

RNA secondary structure cont'd

- Stacks: continuous nested basepairs. (energetically favorable)
- Non-basepaired loops:
  - Hairpin loop.
  - Bulge.
  - Internal loop.
  - Multiloop.



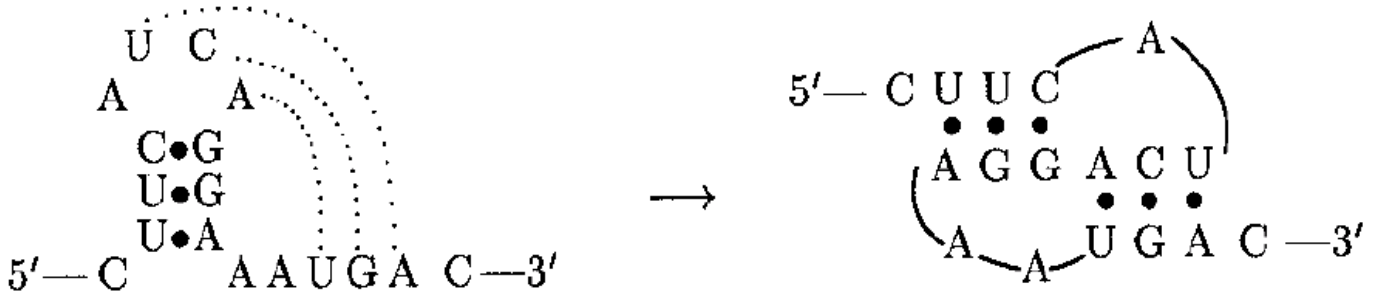
RNA secondary structure cont'd



- Most basepairs are non-crossing basepairs.
  - Any two pairs (*i, j*) and (*i'*,*j'*) ➔ *i* < *i'* < *j'* < *j* or *i'* < *i* < *j* < *j'*
- Pseudoknots are the crossing basepairs.

Pseudoknots

- Pseudoknots are important for certain ncRNAs
- Violate the non-crossing assumption.
- Pseudoknots make most problems harder
- We assume there are no pseudoknots otherwise noted.



[Rivas and Eddy (1999)]

RNA secondary structure prediction

- It is a basic issue in ncRNA analysis
- It is important information to the biologists.
- Searching and alignment algorithms are based on these models.
- RNA secondary structure -- a set of non-crossing base pairs.

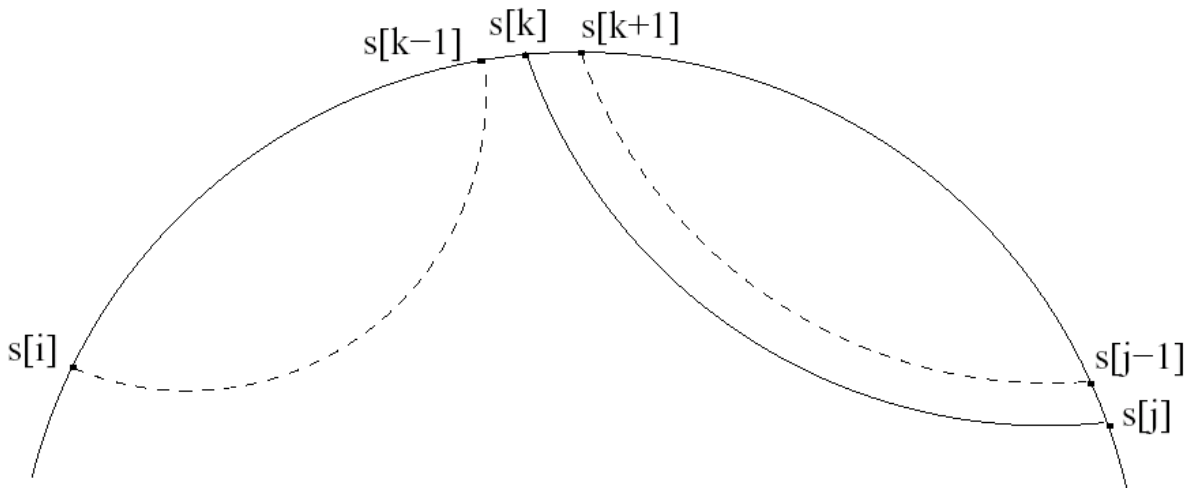
Base pair maximization problem

- 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.

A dynamic programming solution

- Let $s[1 \ldots n]$ 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 $s[i \ldots j]$.



[Nussinov (1980)]

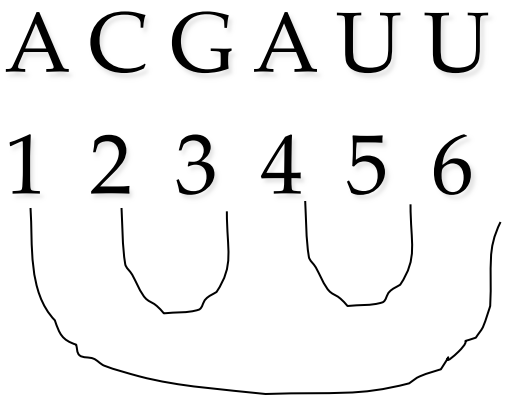
A dynamic programming solution

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

- *M*(*1,n*) is the number of base pairs in the optimal basepaired structure for *s*[*1…n*].
- All these basepairs can be found by tracing back through the matrix *M*.
- Filling *M* needs $O(n^3)$ time.

RNA structure: 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	

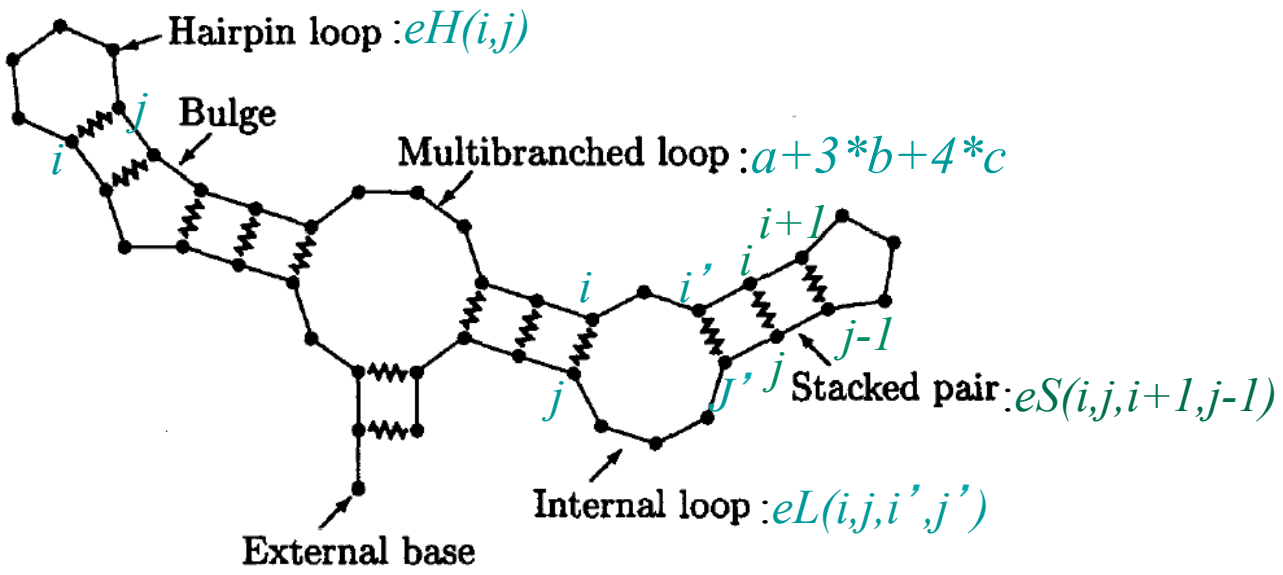


Zuker-Sankoff minimum energy model

- 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 minimum energy model



[Lyngsø (1999)]

RNA minimum energy problem

Problem 2: [RNA minimum energy problem] Given an RNA sequence $s[1...n]$, and four energy functions eH, eS, eL, a, b, and c, determine a noncrossing secondary structure for this RNA sequence such that the sum of the energy over all the loops and stacks in the secondary structure is minimized.

- This problem can be solved by a dynamic programming algorithm in $O(n^4)$ time.
- Lyngsø et al. (1999) revise the energy function for internal loop, proposed an $O(n^3)$ time solution.

Zuker-Sankoff model

- **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.

Recursive functions

- *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.

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

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

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

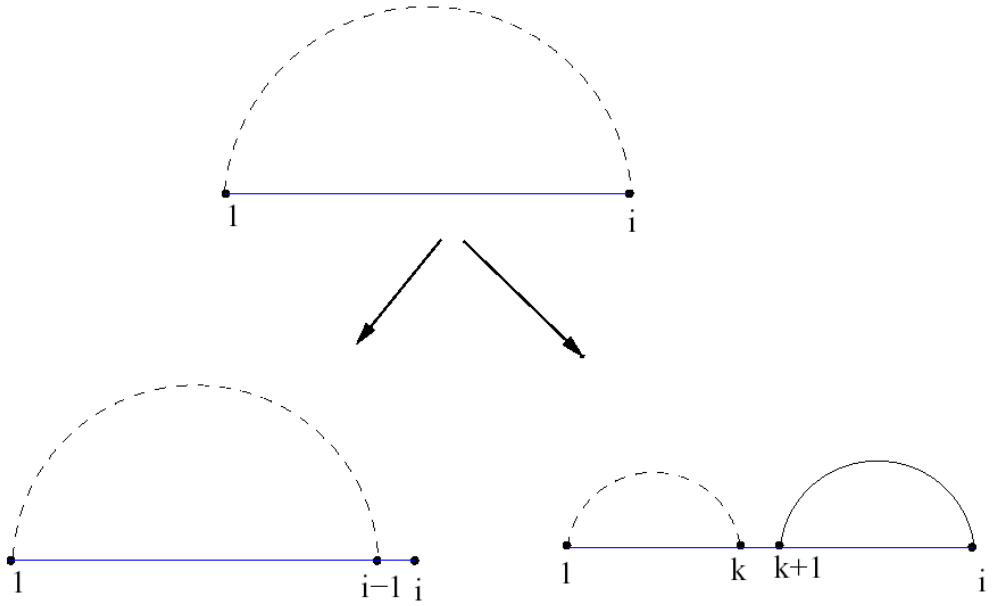
Recursive functions (Zuker)

- *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.

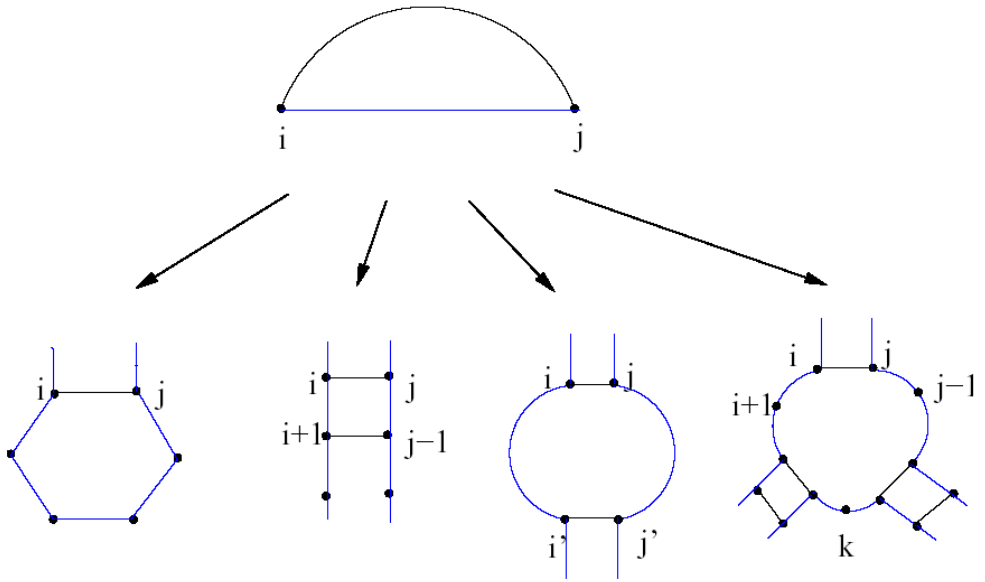
- *WM*(*i,j*) holds the minimum energy of a structure on *s*[*i…j*] that is part of multiloop.



A recursive solution

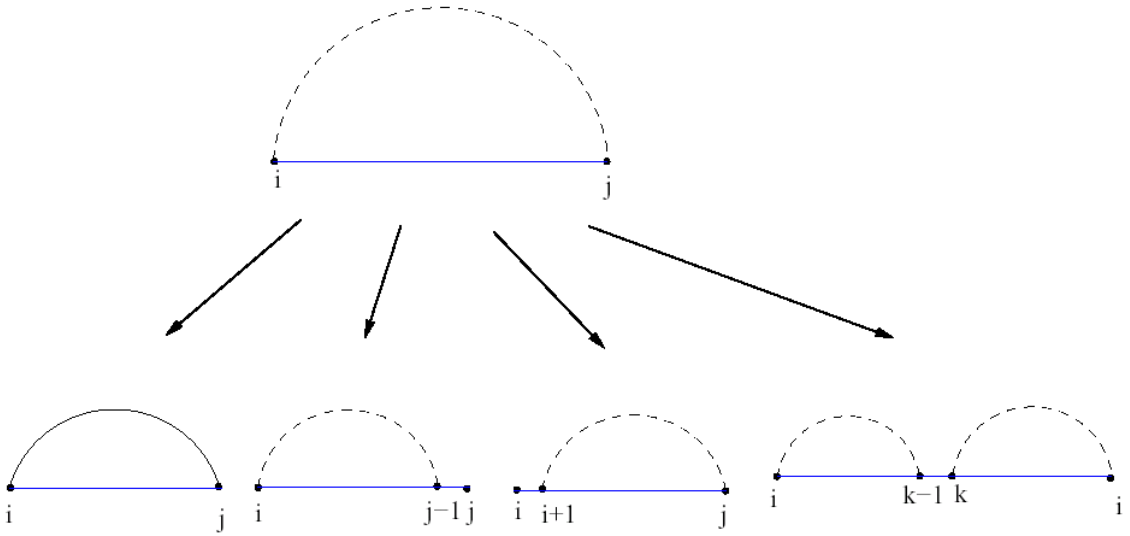
- *W*(*i*) holds the minimum energy of a structure on *s*[*1…i*].
- *V*(*i,j*) holds the minimum energy of a structure on *s*[*i…j*] with *s*[*i*] and *s*[*j*] forming a basepair.
- *WM*(*i,j*) holds the minimum energy of a structure on *s*[*i…j*] that is part of multiloop.

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



A recursive solution

- *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.



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

A recursive solution

- *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.

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

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

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

Prediction with pseudoknots

- Base pair maximization allowing crossing pairs can be solved in polynomial time.

- Ieong et al. (2003) proved that base pairing maximization problem allowing crossing pairs in a *planar* secondary structure is NP-hard.

Prediction with pseudoknots

- Prediction allowing generalized pseudoknots with energy functions depending on adjacent basepairs is NP-hard.
  - Akutsu (2000) (longest common subsequence for multiple sequences (LCS)).
  - Lyngsø and Pedersen (2000) (3SAT).
  - similar to Zuker-Sankoff minimum energy model.
- Pseudoknots in structure-known RNAs.
  - Biologists are not interested in the approximation solutions.
  - Most pseudoknots are planar.
  - Not too many variations.
- Rivas and Eddy (1999) presented a $O(n^6)$ solution allowing most types of pseudoknots in known ncRNAs.