RNA Secondary Structure Prediction

Jianlin Cheng, PhD

School of Electrical Engineering and Computer Science
University of Central Florida

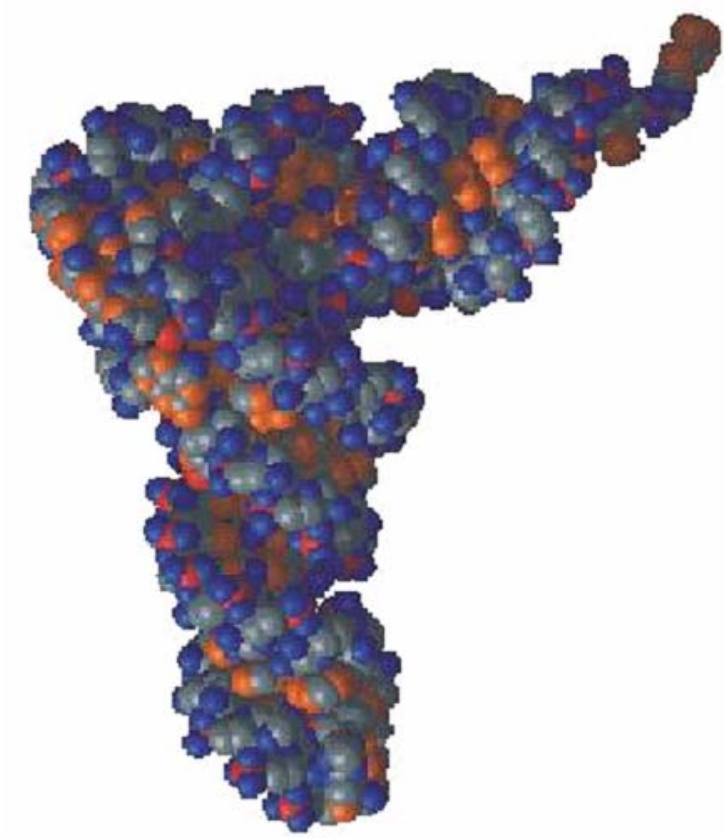
2006

RNA Molecules

- RNA constitutes the bulk of nucleic acid in the cells, being 5-10 times more abundant than DNA

- Best known for its role in transferring genetic information into proteins

- Serves many other important functions in the cell, especially in relation to the regulation of gene expression



Jaco de Ridder, 2004

RNA Function

Types of RNAs	Function of RNA
ribosomal- rRNA	mRNA translation
transfer -tRNA	mRNA translation
messenger -mRNA	protein translation/regulatory
heterogeneous nuclear - hnRNA	intermediates of mRNAs
small cytoplasmic - scRNA	signal recognition particle, tRNA process
small nuclear - snoRNA	mRNA processing, poly A, histone 3' process
small nucleolar- snoRNA	rRNA processing/maturation/methylation
regulatory RNAs	regulation of transcription and translation

L. Samaraweera, 2003

Why Study RNA Structure?

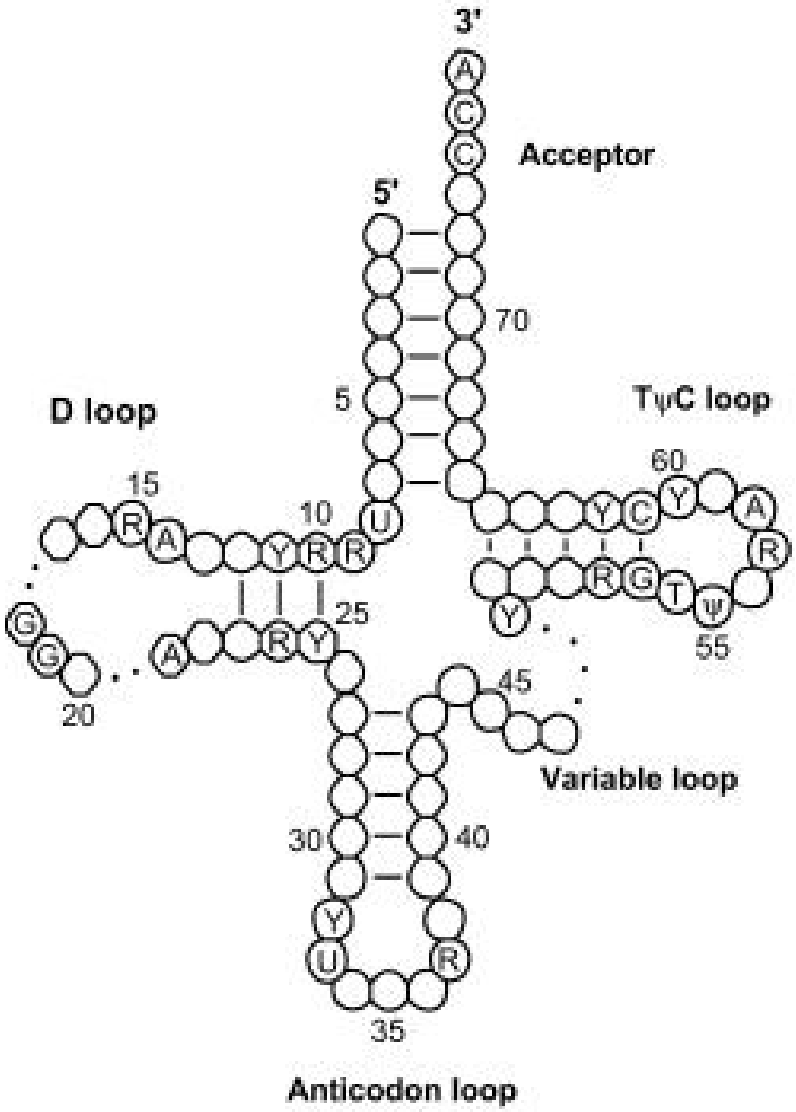
- Major difference between DNA and RNA is that sugar-phosphate backbone part of each nucleotide in DNA lacks an oxygen present on the RNA equivalent
- Difference has a profound effect on the structure and thus potential functions of each type of nucleic acid
- Simple helix structure of DNA effectively limits the range of biological capabilities of DNA
- RNA structure is far more rich and complex, and thus more challenging to solve than that of DNA

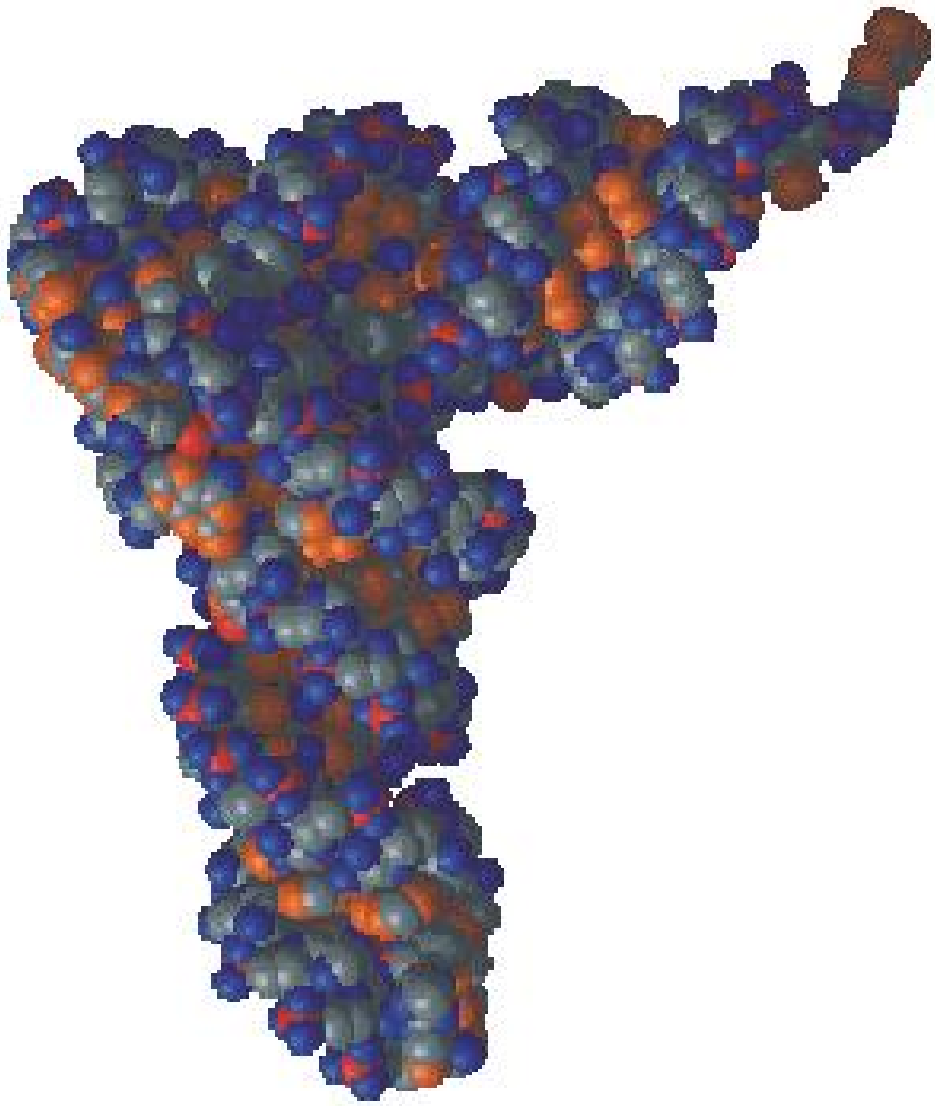
Jaco de Ridder, 2004

Three-Levels of RNA Structure

- Primary structure: sequence
- Secondary structure: intra strand base pairing (Watson-Crick base pairing GC, AU and Wobble base pairing GC) and loops
- Tertiary structure: 3D structure, conformation

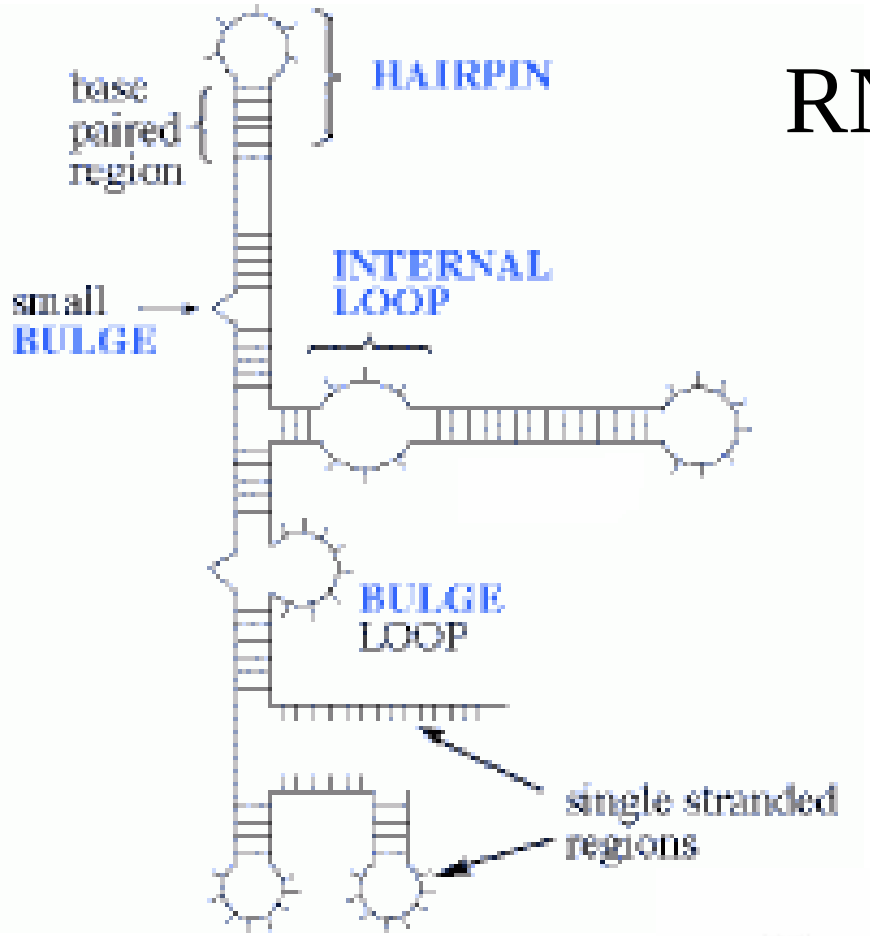
Typical transfer RNA structure



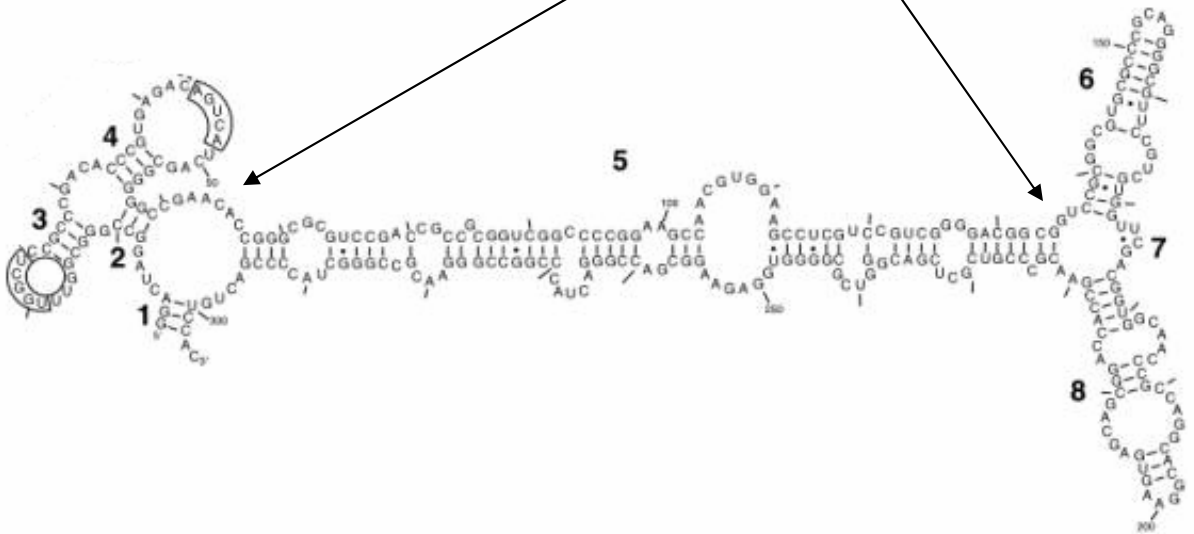


Jaco de Ridder, 2004

RNA Secondary Structure

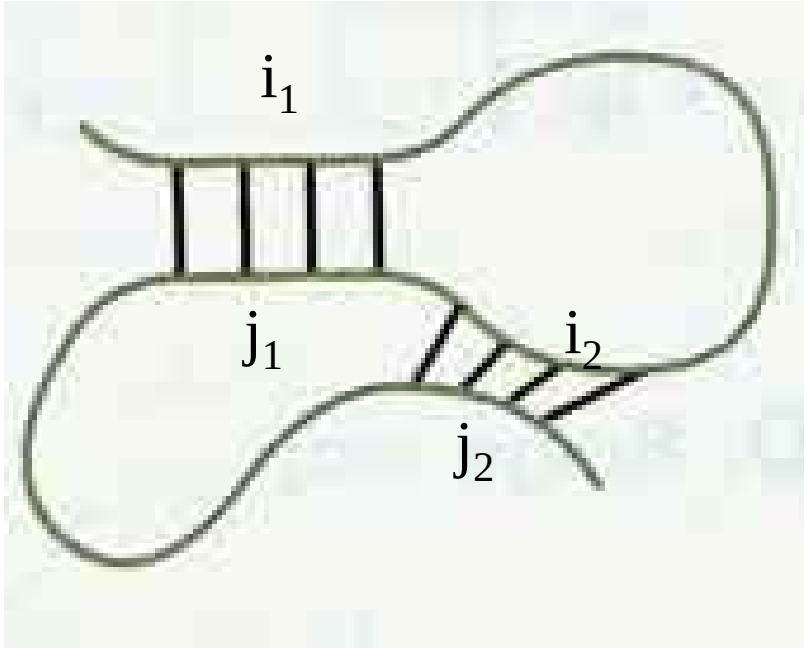


Junction of Multi-Loop



Tertiary structure elements: Pseudoknots

Pseudoknot: interaction of bases inside a loop with bases outside the loop



$$\mathbf{i_1 < i_2 < j_1 < j_2}$$

Sacha Baginsky, 2005

RNA Secondary Structure Stability

Structure stability is dependent upon:

1) The number of GC versus AU and GU base pairs (Higher energy bonds form more stable structures)

2) The number of base pairs in a stem region (longer stems results in more bonds)

3) The number of base pairs in a hairpin loop region (formation of loops with more than 10 or less than 5 bases requires more energy)

4) The number of unpaired bases, whether interior loops or bulges (unpaired bases decrease the stability of the structure)

Jaco de Ridder, 2004

Energy Table for Secondary Structure

- Free-energy values (kcal/mole at 37°C) are as follows:

	Stacking Energies for base pairs					
	A/U	C/G	G/C	U/A	G/U	U/G
A/U	-0.9	-1.8	-2.3	-1.1	-1.1	-0.8
C/G	-1.7	-2.9	-3.4	-2.3	-2.1	-1.4
G/C	-2.1	-2.0	-2.9	-1.8	-1.9	-1.2
U/A	-0.9	-1.7	-2.1	-0.9	-1.0	-0.5
G/U	-0.5	-1.2	-1.4	-0.8	-0.4	-0.2
U/G	-1.0	-1.9	-2.1	-1.1	-1.5	-0.4

	Destabilizing Energies for Loops				
Number of Bases	1	5	10	20	30
Internal	--	5.3	6.6	7.0	7.4
Bulge	3.9	4.8	5.5	6.3	6.7
Hairpin	--	4.4	5.3	6.1	6.5

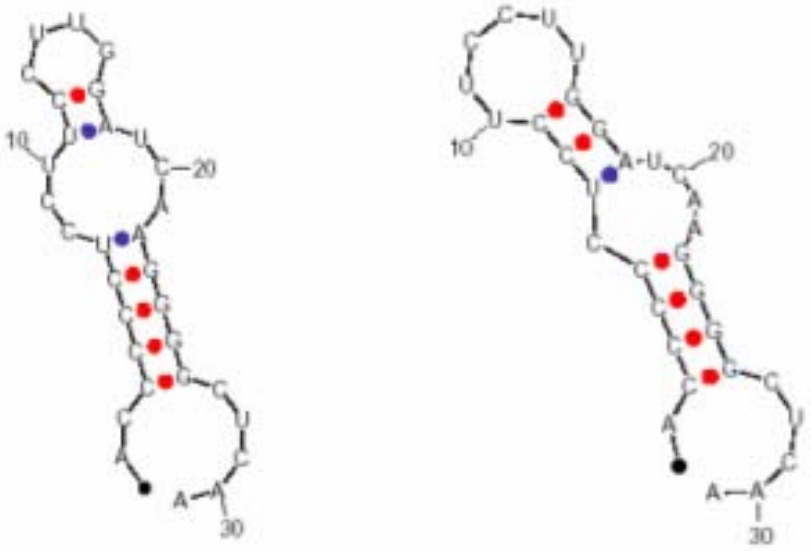
Sacha Baginsky, 2005

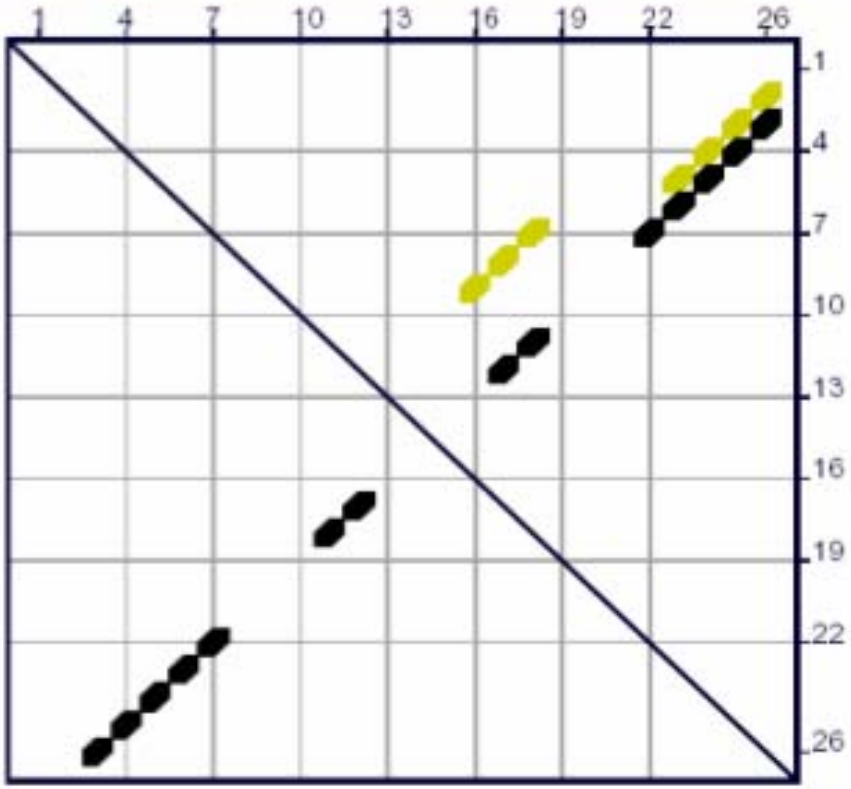
RNA Secondary Structure Prediction Approaches

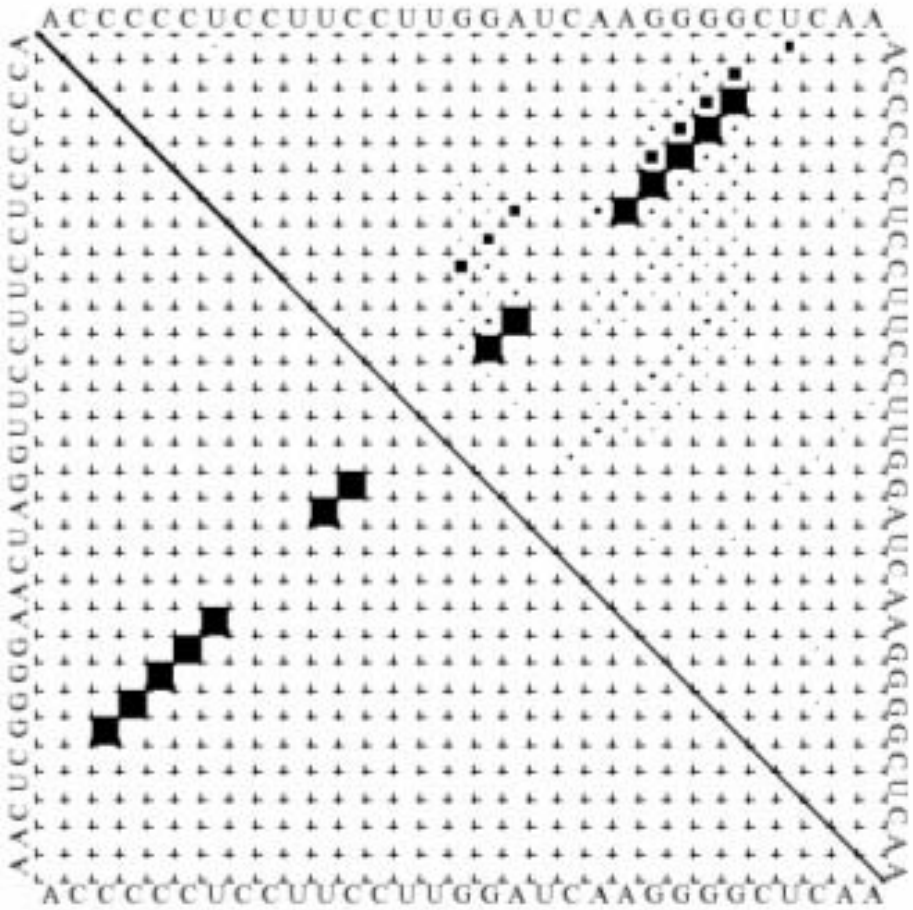
- **Minimum energy**: Look for folds with the lowest free energy (most stable) folds. The fold with more negative free energy, is more stable. The free energy of a fold is the addition of free energy of all motifs found in the structure. Require estimation of energy terms.
- **Comparative Method**: uses multiple sequence alignments of homologous sequences to find conserved regions and covariant base pairs (most trusted if there is enough data)
- Most methods predict secondary structure. Not successful for tertiary structure prediction, which is determined by X-ray and NMR.

Prediction Assumption of Energy-Based Method

- The most likely structure is similar to the energetically most stable structure
- Energy associated with any position in the structure is only influenced by local sequence and structure (previous pair, not next pair)
- No knots.







Jaco de Ridder, 2004

Dynamic programming algorithm

• Recursive definition of the optimal score (We use a simplified scoring function.)

•Initialization of optimal scoring matrix

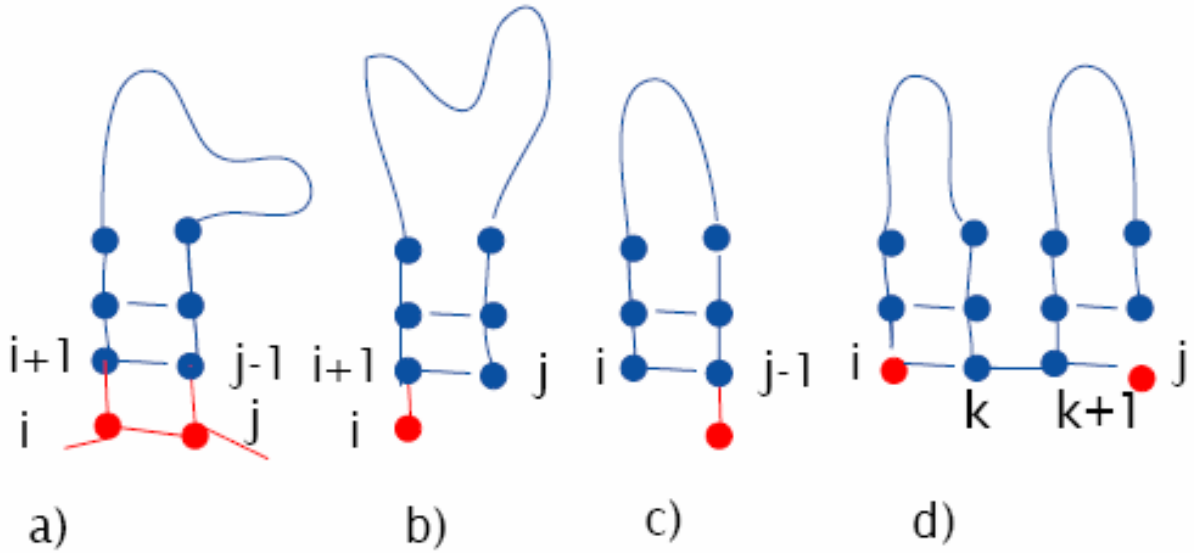
• Bottom-up approach to fill the scoring matrix (bottom-up because smallest subproblems are solved first). Run from diagonal to diagonal

• Traceback of the matrix to recover the global, optimal solution

Dynamic programming approach

Let $E(i,j)$ = minimum energy for subchain starting at i and ending at j

$\alpha(ri,rj)$ = energy of pair ri, rj (rj = base at position j)



a) i,j is paired $E(i,j) = E(i+1,j-1) + \alpha(ri,rj)$

b) i is unpaired $E(i,j) = E(i+1,j)$

c) j is unpaired $E(i,j) = E(i,j-1)$

d) bifurcation $E(i,j) = E(i,k)+E(k+1,j)$

Sacha Baginsky, 2005

Dynamic Programming Algorithm for RNA Secondary Structure Prediction

- Given a RNA sequence: $x_1,x_2,x_3,\ldots,x_L$ and a scoring function a(x,y)
- **Initialization**: E[i,i-1] = 0, E[i,i] = 0
- **Recursion**:

for d = 1,2,3,4,…,L-1
{
    for (i = 1; i + d <= L; i++)
    {
        j = i + d;
        E[i,j] = min { E[i+1,j],
                        E[i,j-1],
                        E[i+1,j-1] + a(r_i,r_j)
                        $\min_{i<k<j}$ (E[i,k] + E[k+1,j])
                    }
    }
}

Note: i is always smaller than j.

Dynamic Programming Algorithm for RNA Secondary Structure Prediction

- Given a RNA sequence: $x_1,x_2,x_3,\ldots,x_L$ and a scoring function a(x,y)
- **Initialization**: E[i,i-1] = 0, E[i,i] = 0
- **Recursion**:

```
for d = 1,2,3,4,…,L-1
{
        for ( i = 1; i + d <= L; i++)
        {
                j = i + d;
                E[i,j] = min {      E[i+1,j],
                                    E[i,j-1],
                                    E[i+1,j-1] + a(r_i,r_j)
                                    min_{i<k<j} ( E[i,k] + E[k+1,j] )
                            }
        }
}
```

Time Complexity?

Note: i is always smaller than j.

A Simple Example

Input: GGAAAUCC

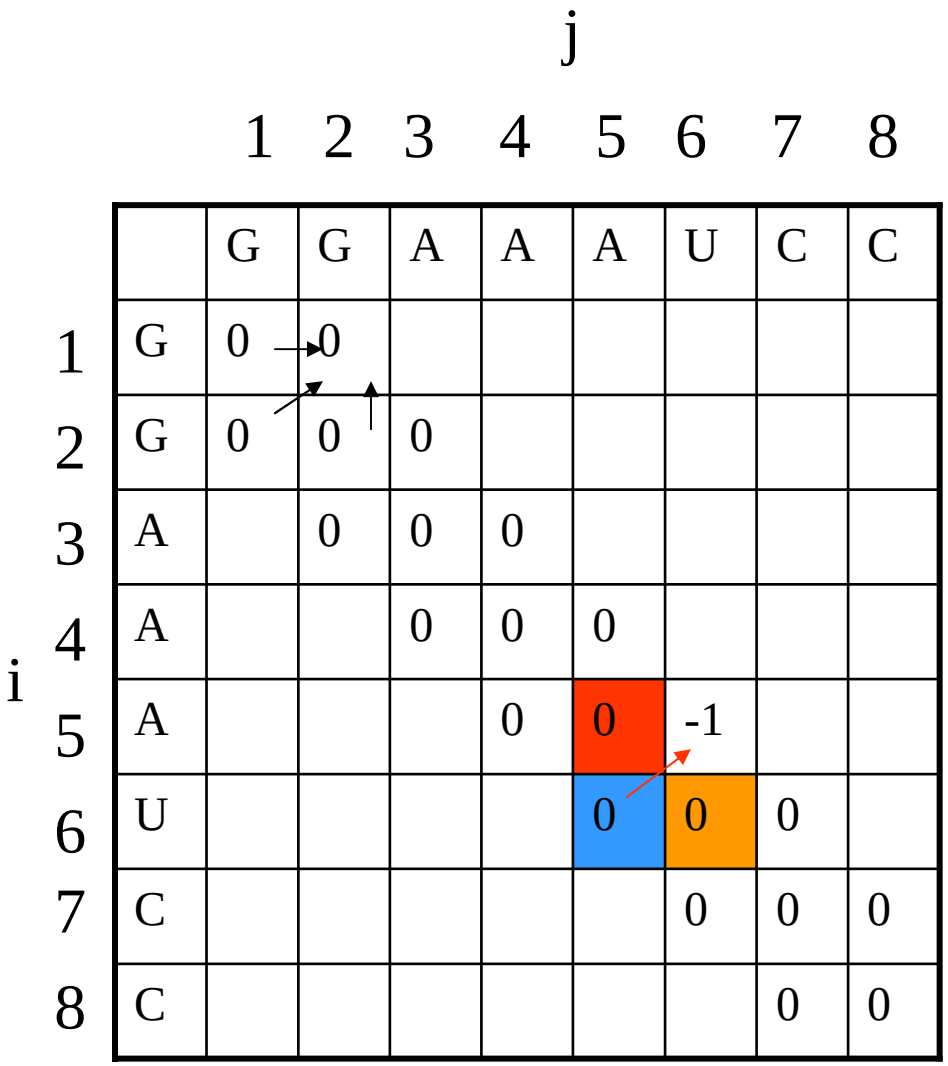
Scoring Function: $a(r_i,r_j) = -1$ if r_i and r_j form a Watson-Crick base pair. Otherwise, 0.

i \ j		1 G	2 G	3 A	4 A	5 A	6 U	7 C	8 C
1	G	0							
2	G	0	0						
3	A		0	0					
4	A			0	0				
5	A				0	0			
6	U					0	0		
7	C						0	0	
8	C							0	0

Initialization

$E[i,i-1] = 0$, $E[i,i] = 0$

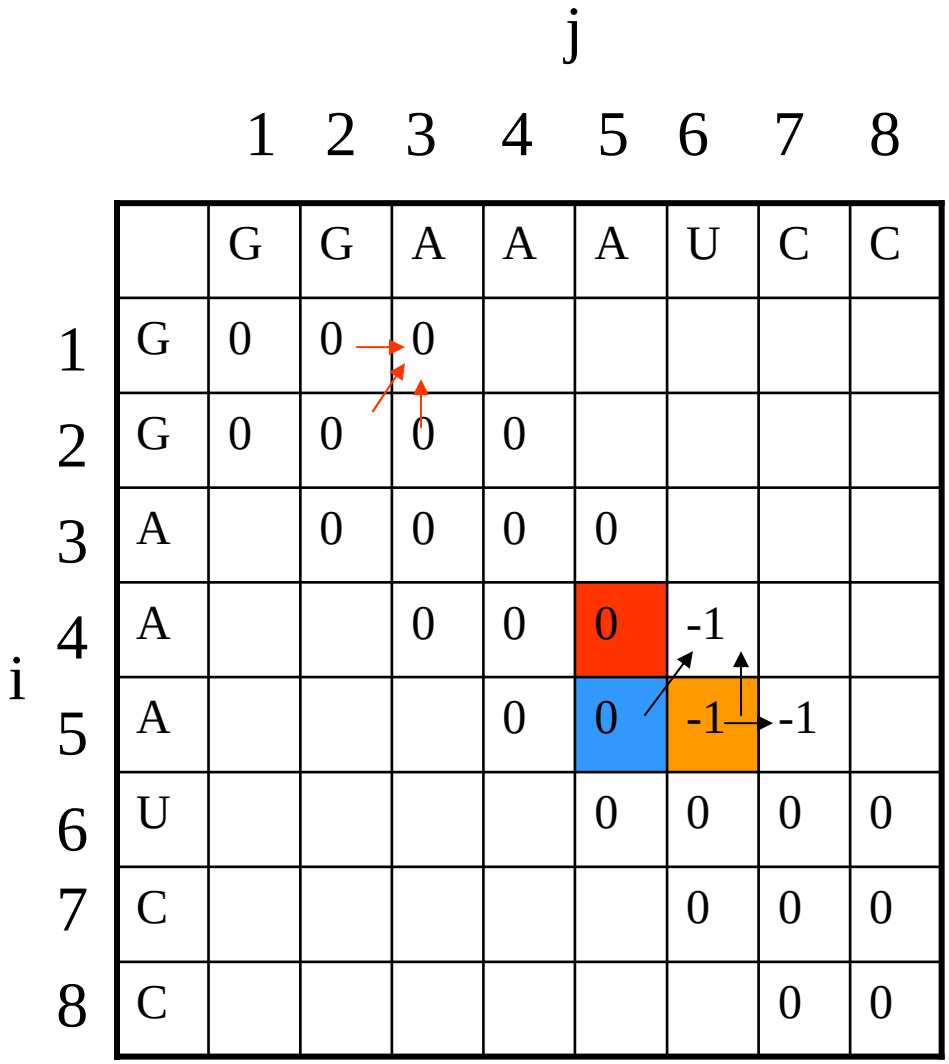
Fill matrix from diagonal to diagonal



		G	G	A	A	A	U	C	C
1	G	0	0						
2	G	0	0	0					
3	A		0	0	0				
4	A			0	0	0			
5	A				0	0	-1		
6	U					0	0	0	
7	C						0	0	0
8	C							0	0

E[1,2] = min(
E(1,2-1),
E(1+1,2),
E[1+1,2-1]+a(1,2)
~~$\min_k(E[1,k]+E[k+1,2]$~~
)
= 0

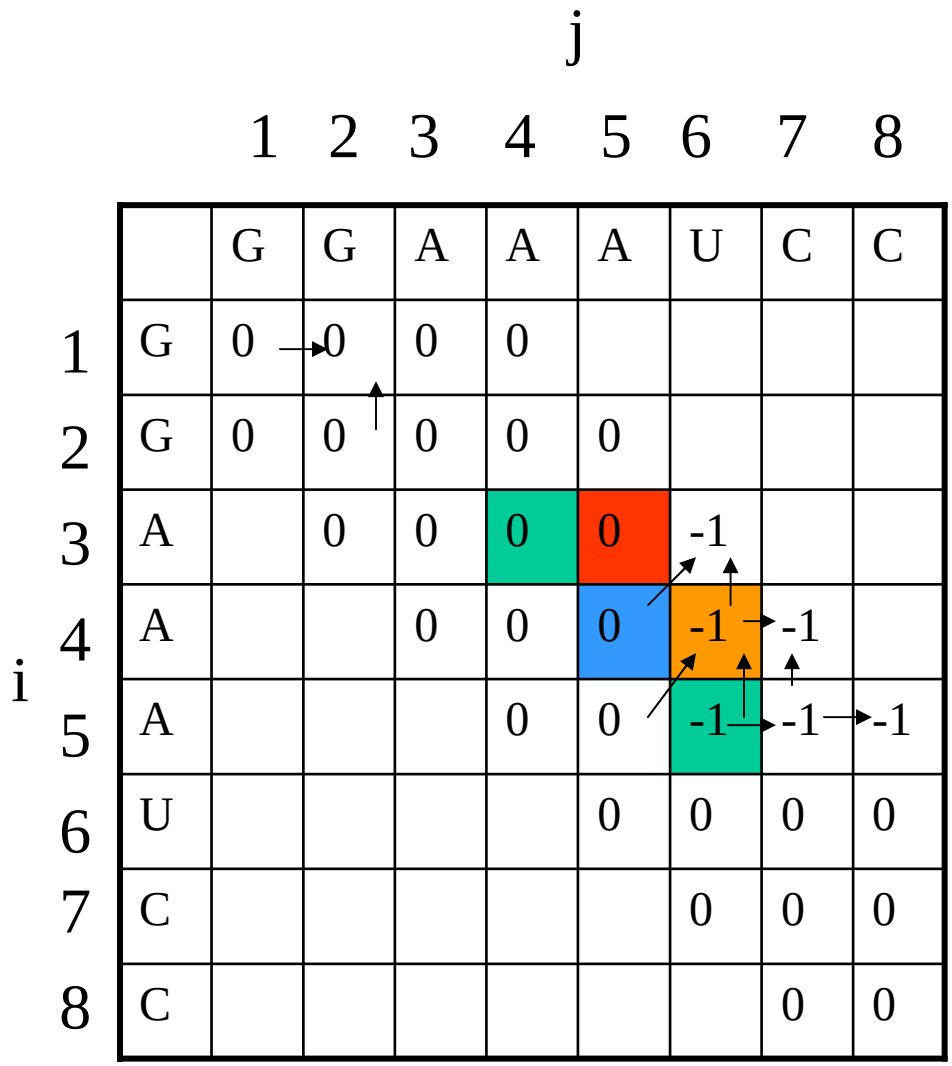
E[5,6] = min(
E[5,5],
E[6,5]+a(5,6),
E[6,6])
~~$\min_k(E[5,k]+E[k,6])$~~
)
= -1



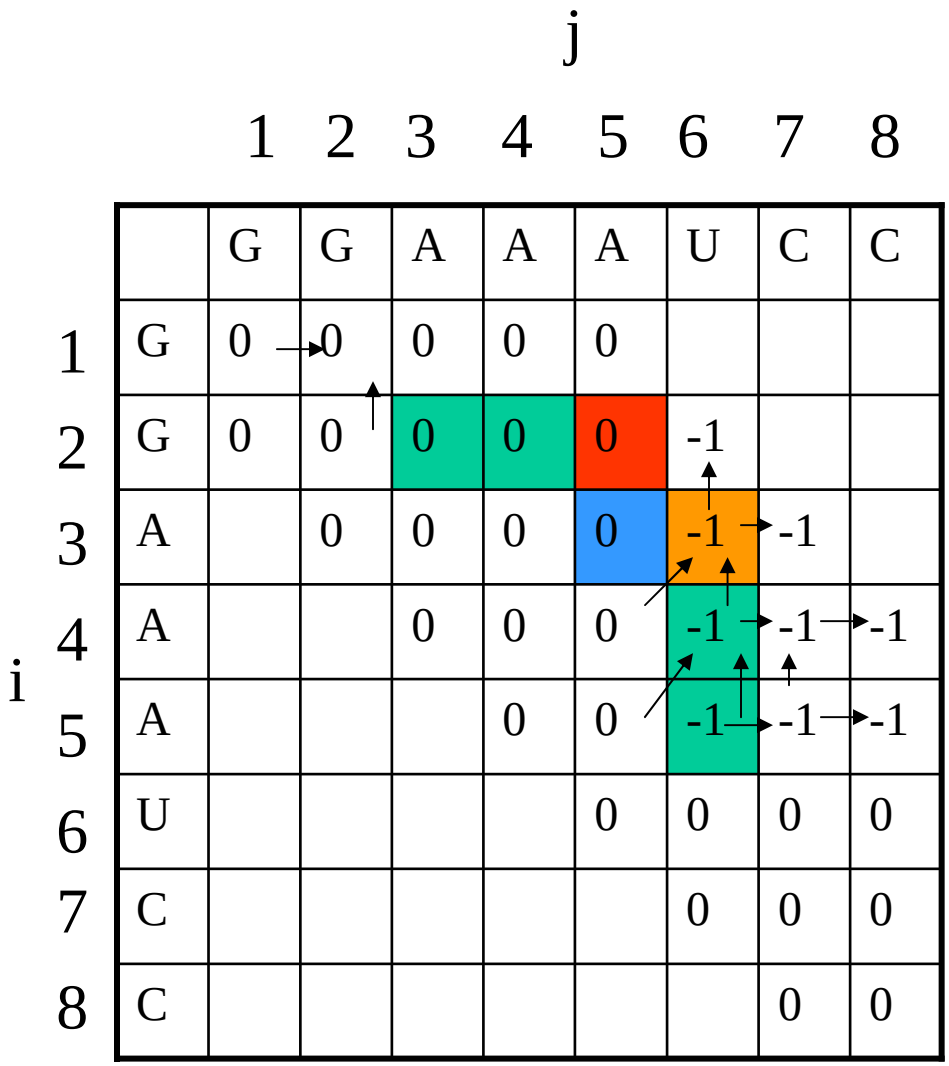
E[1,3] = min(
E(1,2),
E(2,3),
E[2,2] + a(G,A)
)
= 0

Any valid k?

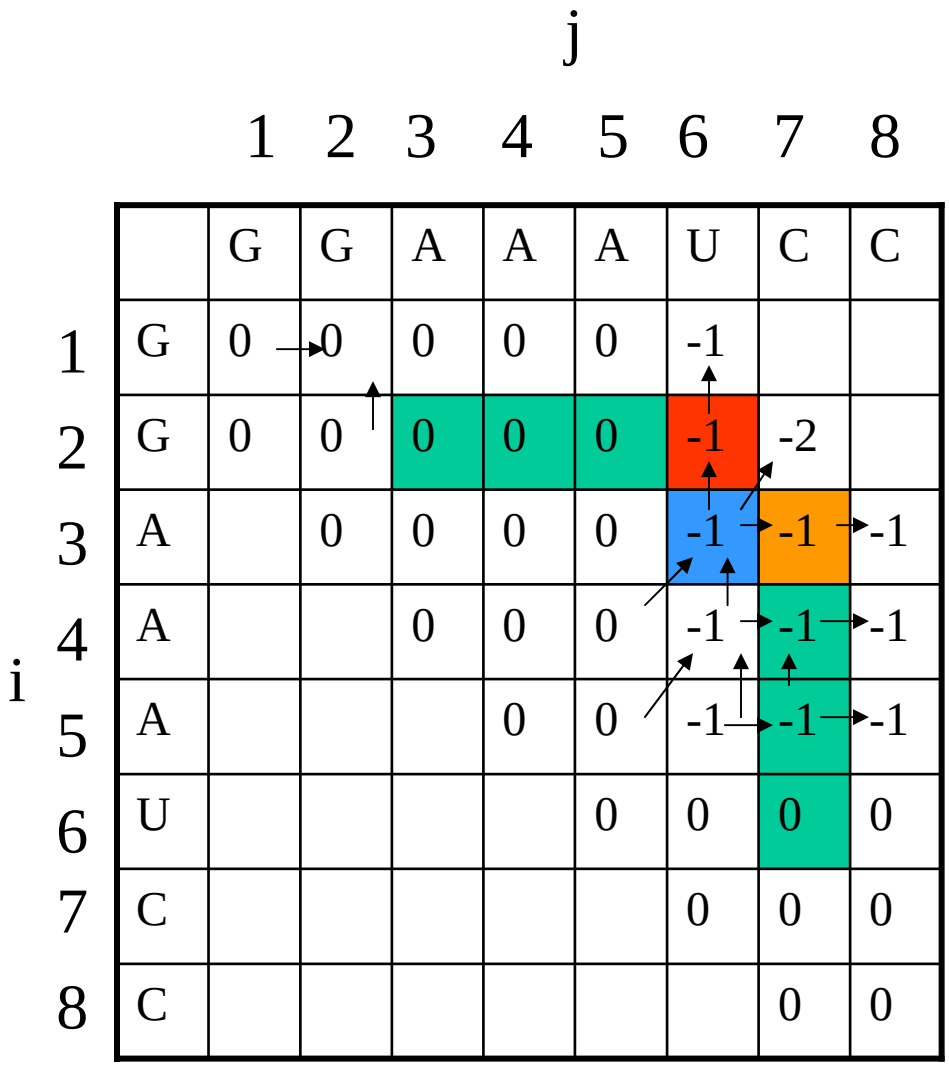
E[4,6] = min (
E[4,5],
E[5,6],
E[5,5] + a(A,U)
)
= -1



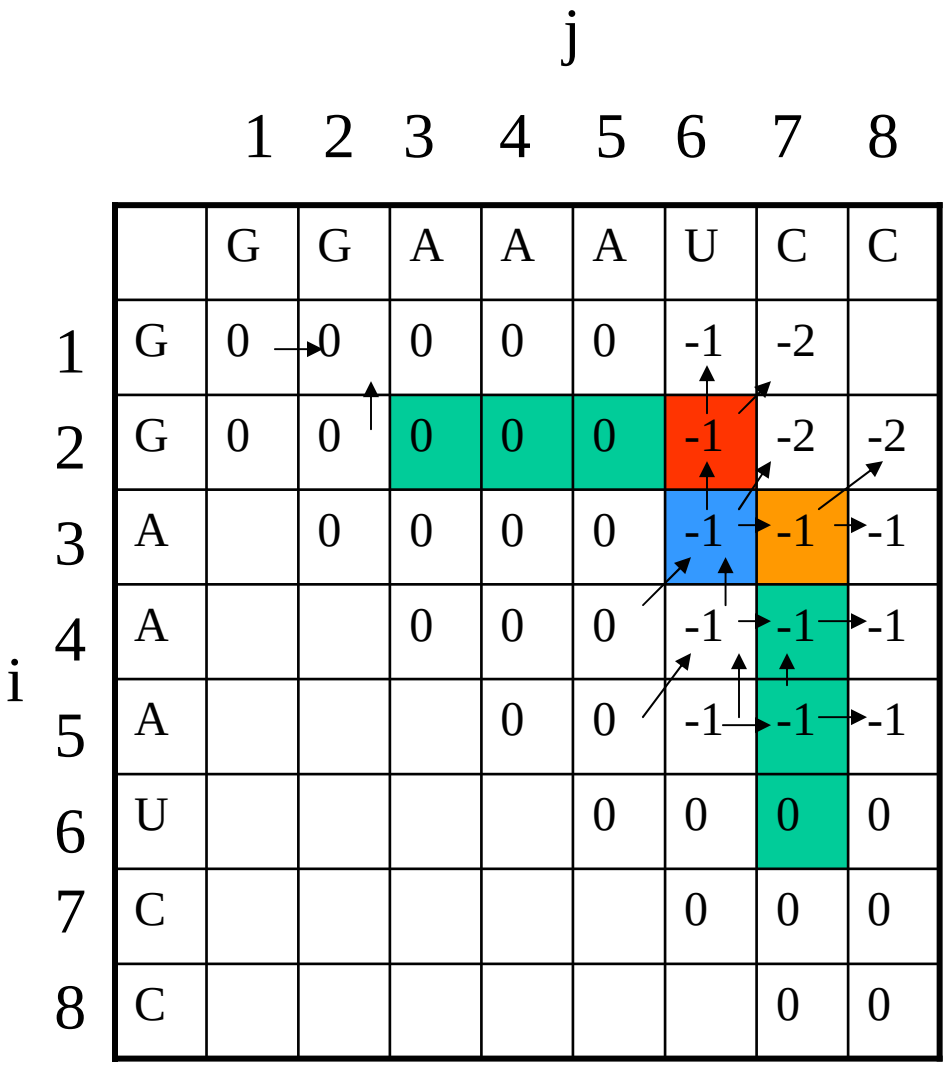
E[3,6] = min(
E[3,5],
E[4,6],
E[4,5]+a(A,U),
E[3,4]+E[5,6]
)
= -1



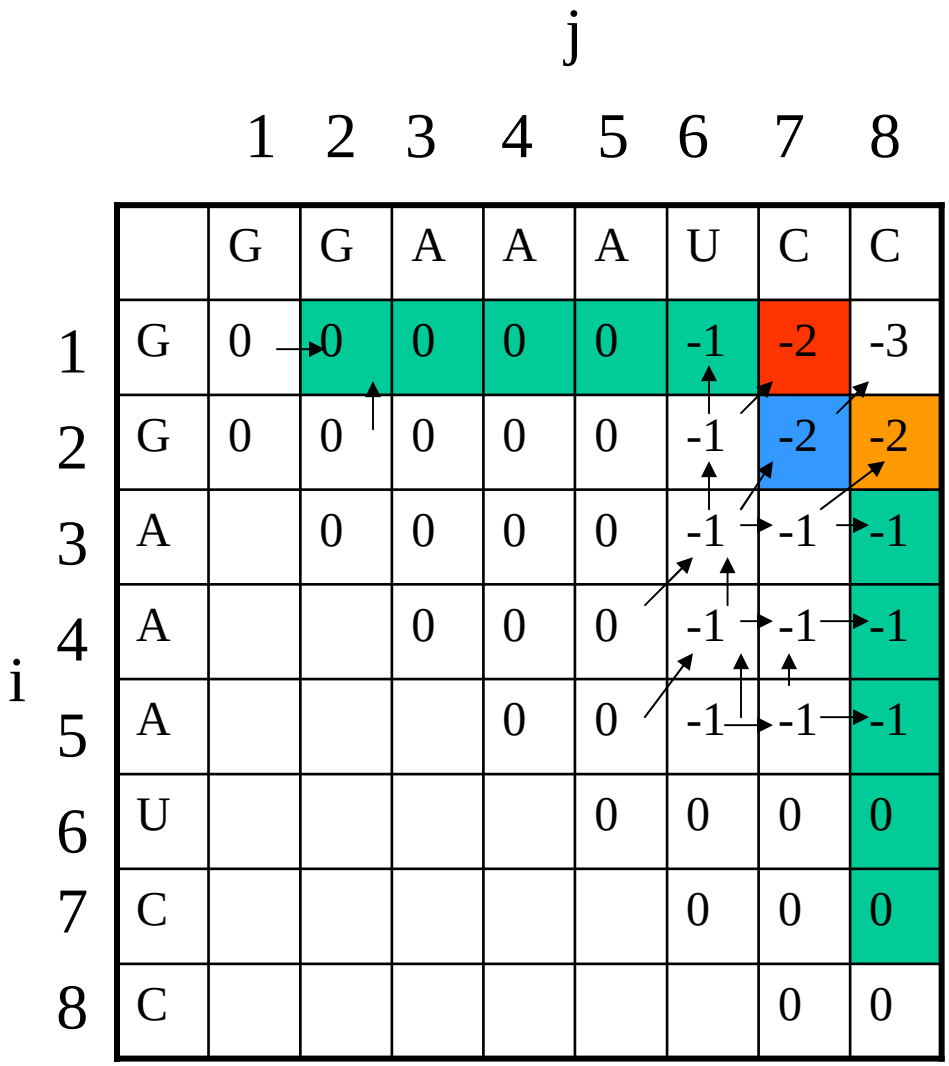
E[2,6] = min(
E[2,5],
E[3,6],
E[3,5]+a(G,U),
E[2,3]+E[4,6],
E[2,4]+E[5,6]
) = -1



E[2,7] = min (
E[2,6],
E[3,7],
E[3,6]+a(G,C),
E[2,3] + E[4,7],
E[2,4] + E[5,7],
E[2,5] + E[6,7]
)
= -2



E[1,7] = min(
E(1,6),
E(2,7),
E[2,6] + a(G,C),
E[1,2] + E[3,7],
E[1,3] + E[4,7],
E[1,4] + E[5,7],
E[1,5] + E[6,7]
)
= -2



Best score:

$E[1,8] = E[2,7] + a(G,C)$

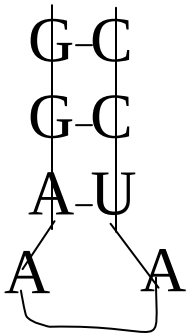
$= -3$

Time Complexity?

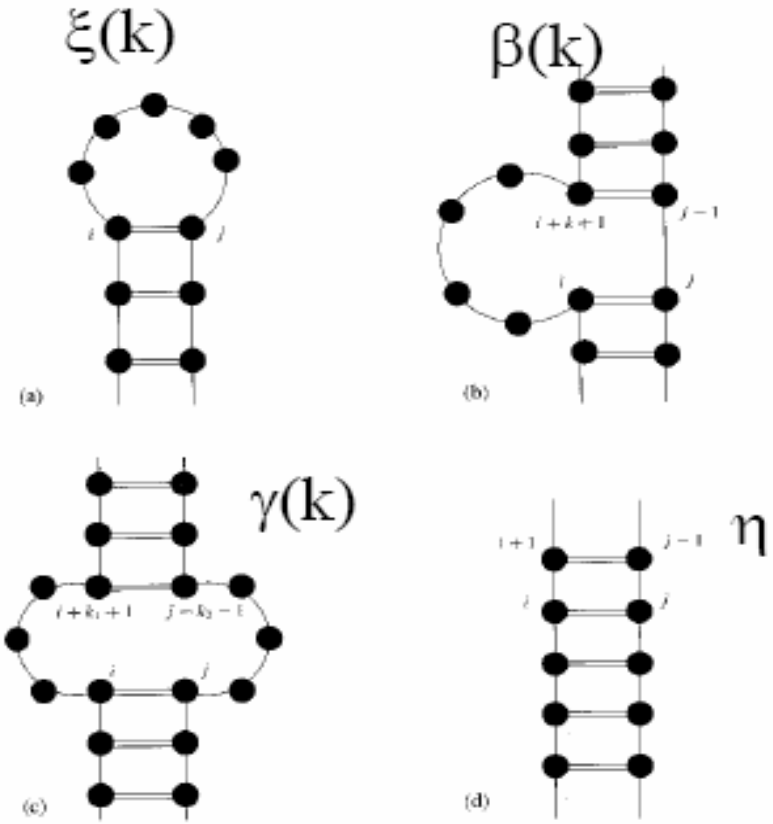
Trace Back

i \ j		1 G	2 G	3 A	4 A	5 A	6 U	7 C	8 C
1	G	0	0	0	0	0	-1	-2	-3
2	G	0	0	0	0	0	-1	-2	-2
3	A		0	0	0	0	-1	-1	-1
4	A			0	0	0	-1	-1	-1
5	A				0	0	-1	-1	-1
6	U					0	0	0	0
7	C						0	0	0
8	C							0	0

Best score: E[1,8] = -3



Even more realistic energy function



Loops have destabilizing effect structure (d) should have lower energy that (b).

Destabilizing contribution of loops should depend on the loop length (k).

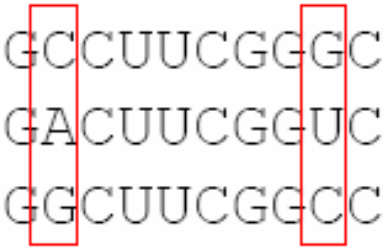
Stacking has additional stabilizing contribution η.

So in reality, more realistic energy function that considers different loops are needed. But the basic idea of dynamic programming is still applied.

Sacha Baginsky, 2005

Covariance method

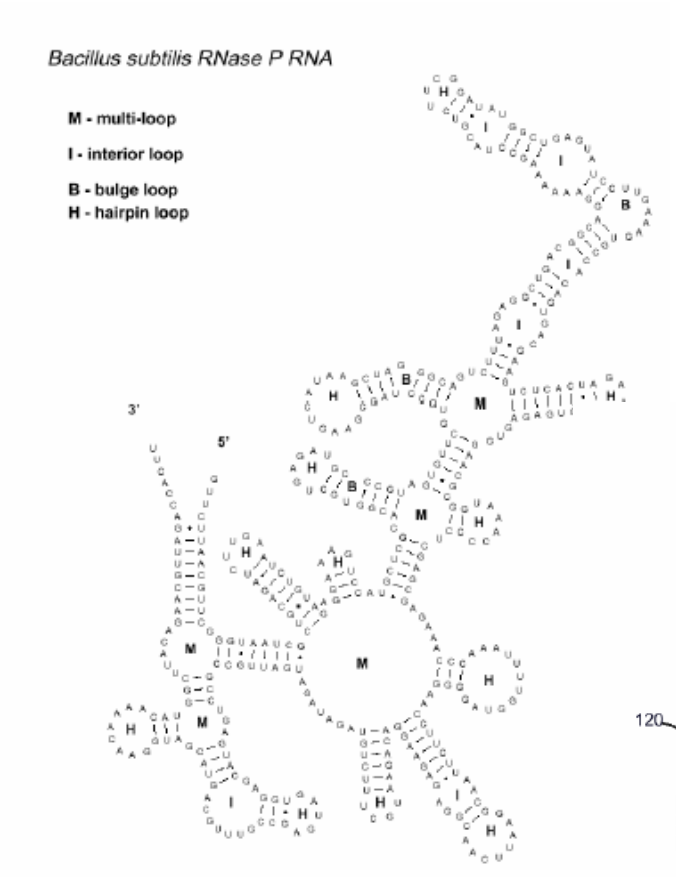
In a correct multiple alignment RNAs, conserved base pairs are often revealed by the presence of frequent correlated compensatory mutations.

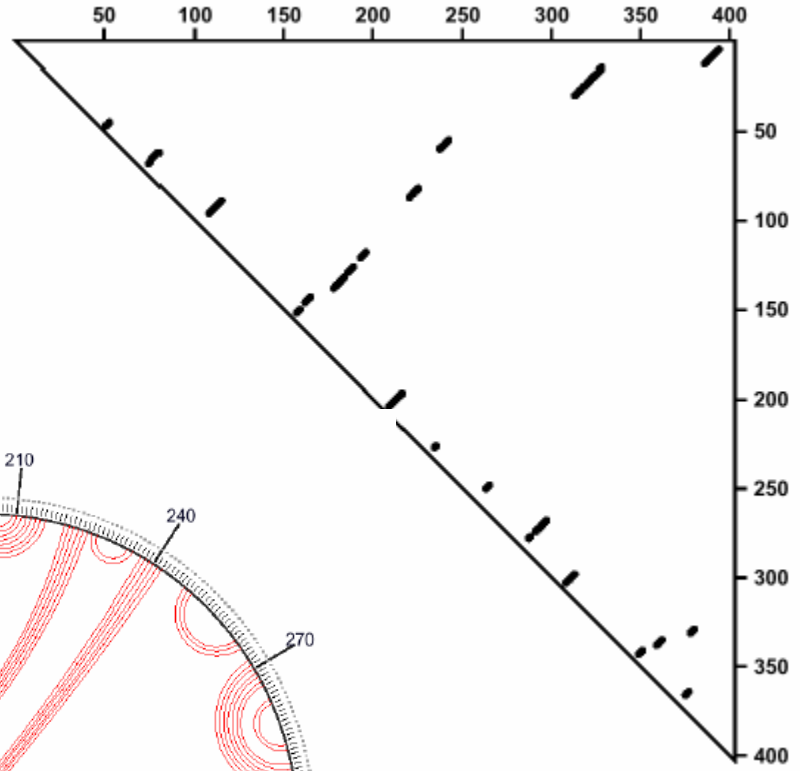


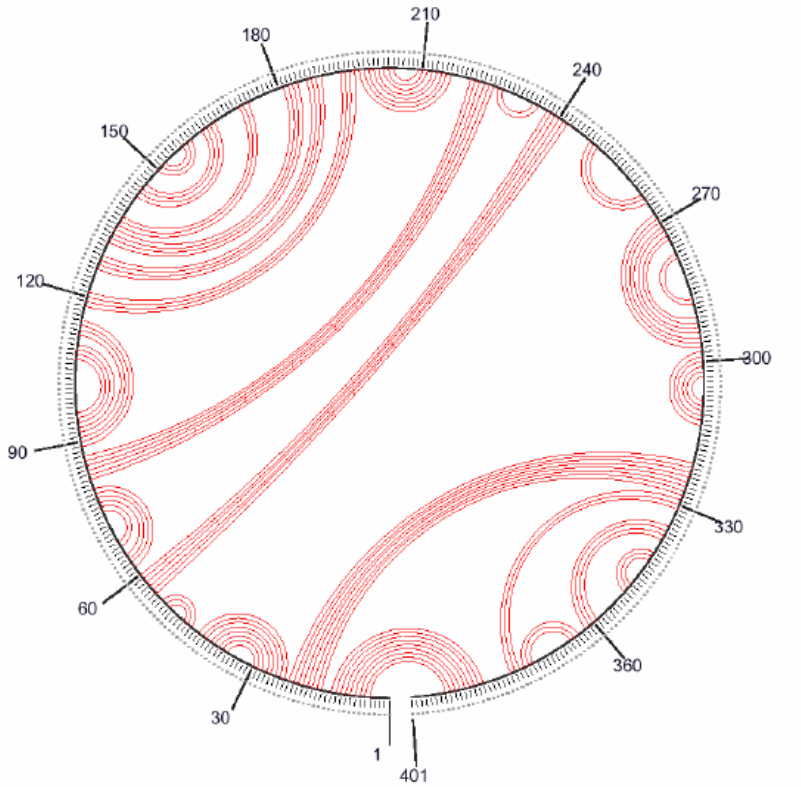
Two boxed positions are covarying to maintain Watson-Crick complementary. This covariation implies a base pair which may then be extended in both directions.

More information: www.rna.icmb.utexas.edu/METHODS/menu.html

Representation of RNA secondary structures

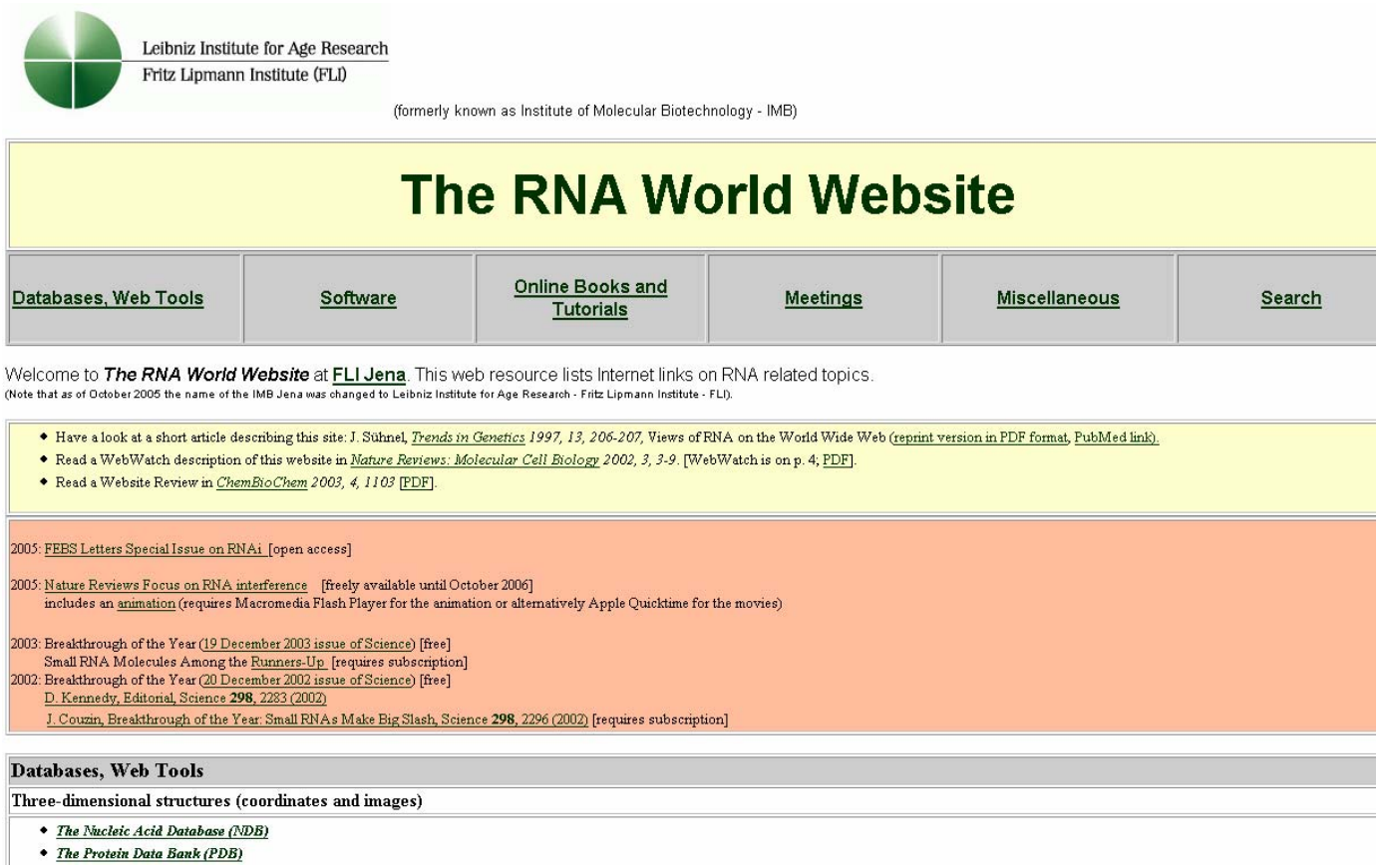






RNA Resources

Web: http://www.imb-jena.de/RNA.html



Leibniz Institute for Age Research
Fritz Lipmann Institute (FLI)

(formerly known as Institute of Molecular Biotechnology - IMB)

The RNA World Website

Databases, Web Tools	Software	Online Books and Tutorials	Meetings	Miscellaneous	Search

Welcome to ***The RNA World Website*** at FLI Jena. This web resource lists Internet links on RNA related topics.
(Note that as of October 2005 the name of the IMB Jena was changed to Leibniz Institute for Age Research - Fritz Lipmann Institute - FLI).

- Have a look at a short article describing this site: J. Sühnel, *Trends in Genetics 1997, 13, 206-207*, Views of RNA on the World Wide Web (reprint version in PDF format, PubMed link).
- Read a WebWatch description of this website in *Nature Reviews: Molecular Cell Biology 2002, 3, 3-9*. [WebWatch is on p. 4; PDF].
- Read a Website Review in *ChemBioChem 2003, 4, 1103* [PDF].

2005: FEBS Letters Special Issue on RNAi [open access]

2005: Nature Reviews Focus on RNA interference [freely available until October 2006]
includes an animation (requires Macromedia Flash Player for the animation or alternatively Apple Quicktime for the movies)

2003: Breakthrough of the Year (19 December 2003 issue of Science) [free]
Small RNA Molecules Among the Runners-Up [requires subscription]
2002: Breakthrough of the Year (20 December 2002 issue of Science) [free]
D. Kennedy, Editorial, Science **298**, 2283 (2002)
J. Couzin, Breakthrough of the Year: Small RNAs Make Big Slash, Science **298**, 2296 (2002) [requires subscription]

Databases, Web Tools

Three-dimensional structures (coordinates and images)

- ***The Nucleic Acid Database (NDB)***
- ***The Protein Data Bank (PDB)***

RNA Folding Software

- Vienna: http://www.tbi.univie.ac.at/RNA/
- MFold: http://www.bioinfo.rpi.edu/applications/mfold/rna/form1.cgi
- AliFold: http://rna.tbi.univie.ac.at/cgi-bin/alifold.cgi (use aligned sequences)
- Genebee: http://www.genebee.msu.su/services/rna2_reduced.html (alignment)

Ten Topics

- 1. Introduction to Molecular Biology and Bioinformatics
- 2. Pairwise Sequence Alignment Using Dynamic Programming
- 3. Practical Sequence/Profile Alignment Using Fast Heuristic Methods (BLAST and PSI-BLAST)
- 4. Multiple Sequence Alignment
- 5. Gene Identification
- 6. Phylogenetic Analysis
- 7. Protein Structure Analysis and Prediction
- **8. RNA Secondary Structure Prediction**
- 9. Clustering and Classification of Gene Expression Data
- 10. Search and Mining of Biological Databases, Databanks, and Literature