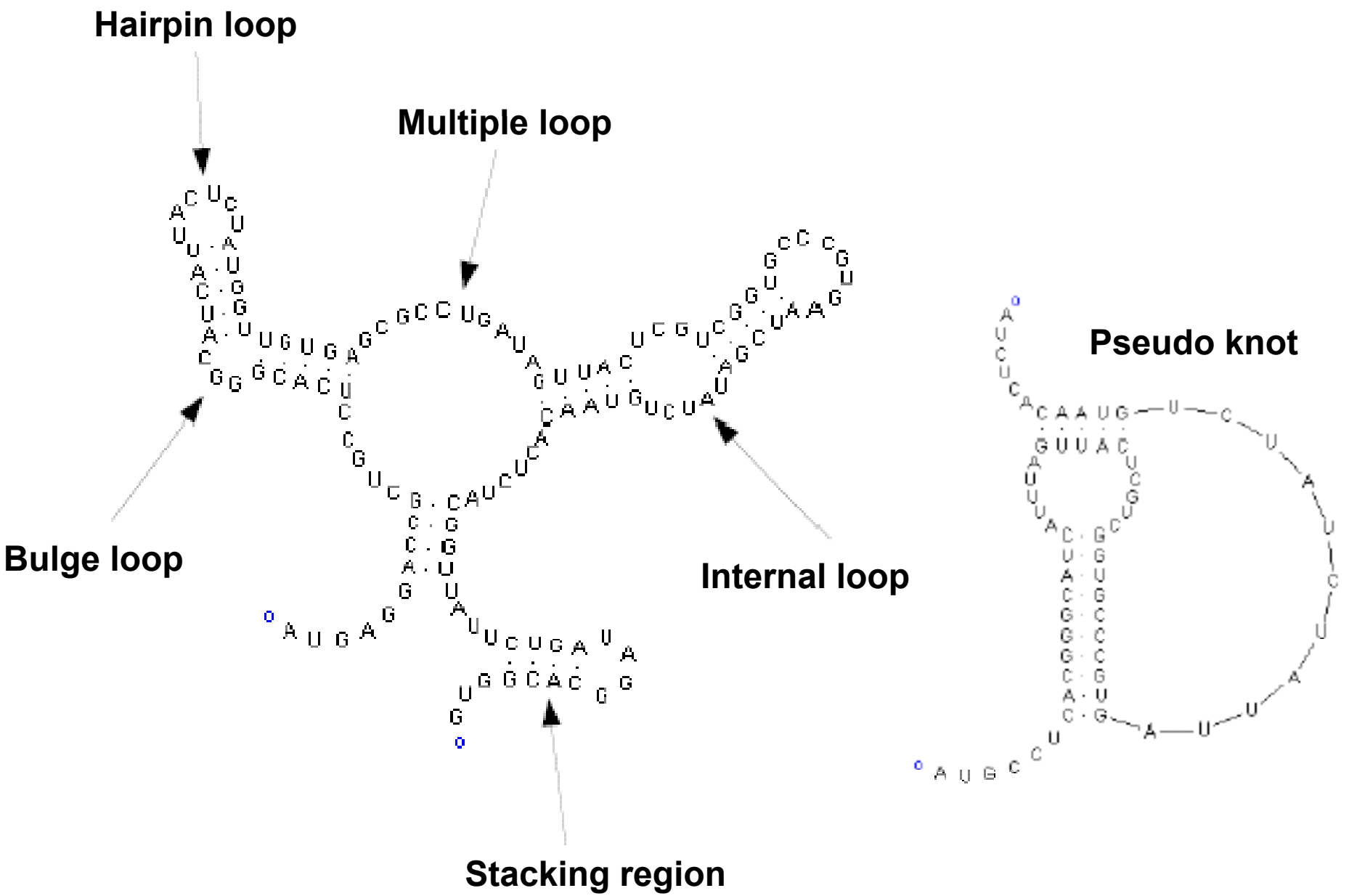
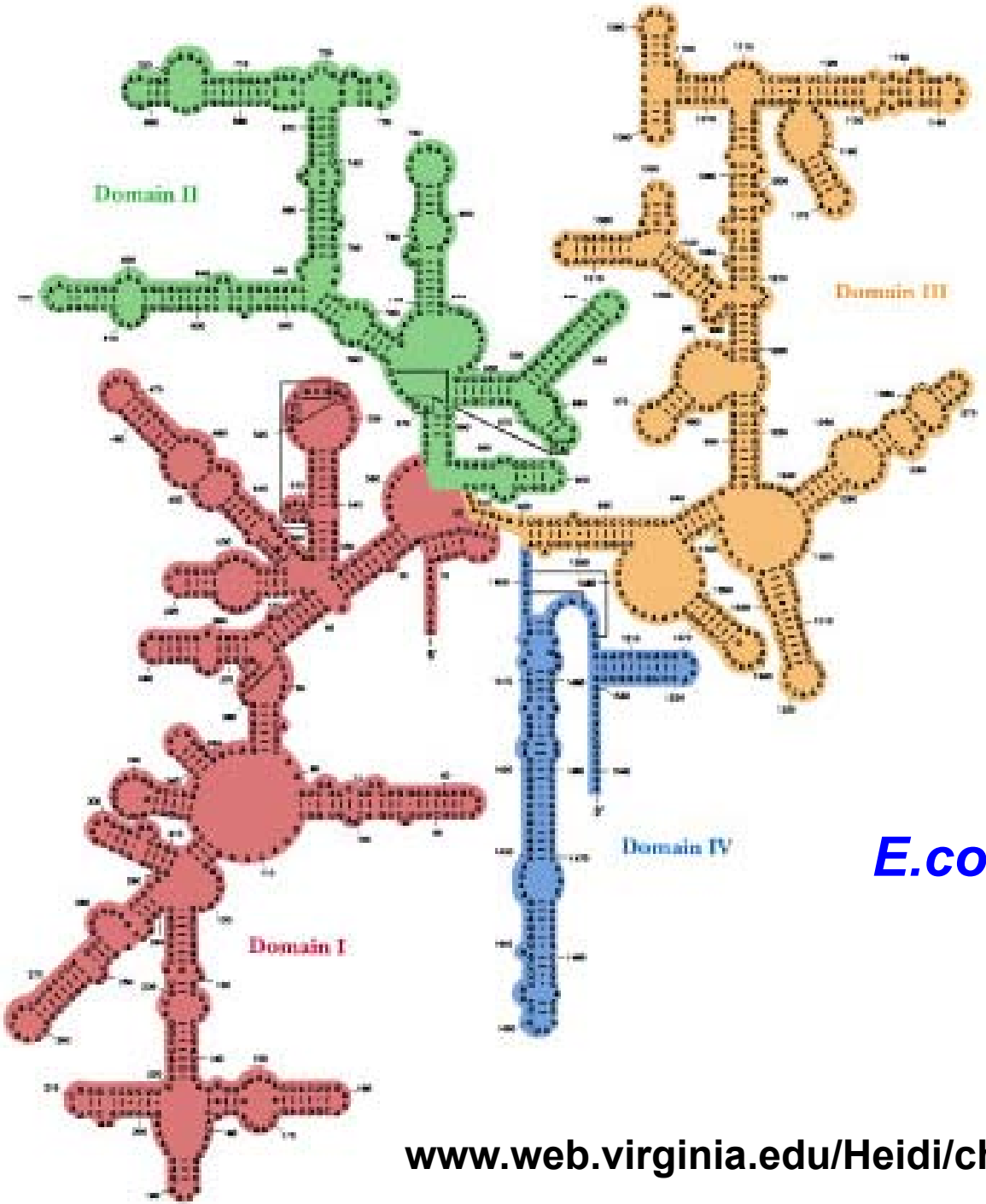
#11
RNA Secondary Structure Prediction

Topics:

- RNA Secondary Structure
  - Hairpin/Bulge/Internal/Multi loops, Stacking region
  - Pseudoknot
- Nussinov algorithm
  - Initialization, Matrix fill stage, Traceback stage
- Some Expansions
- Zuker method

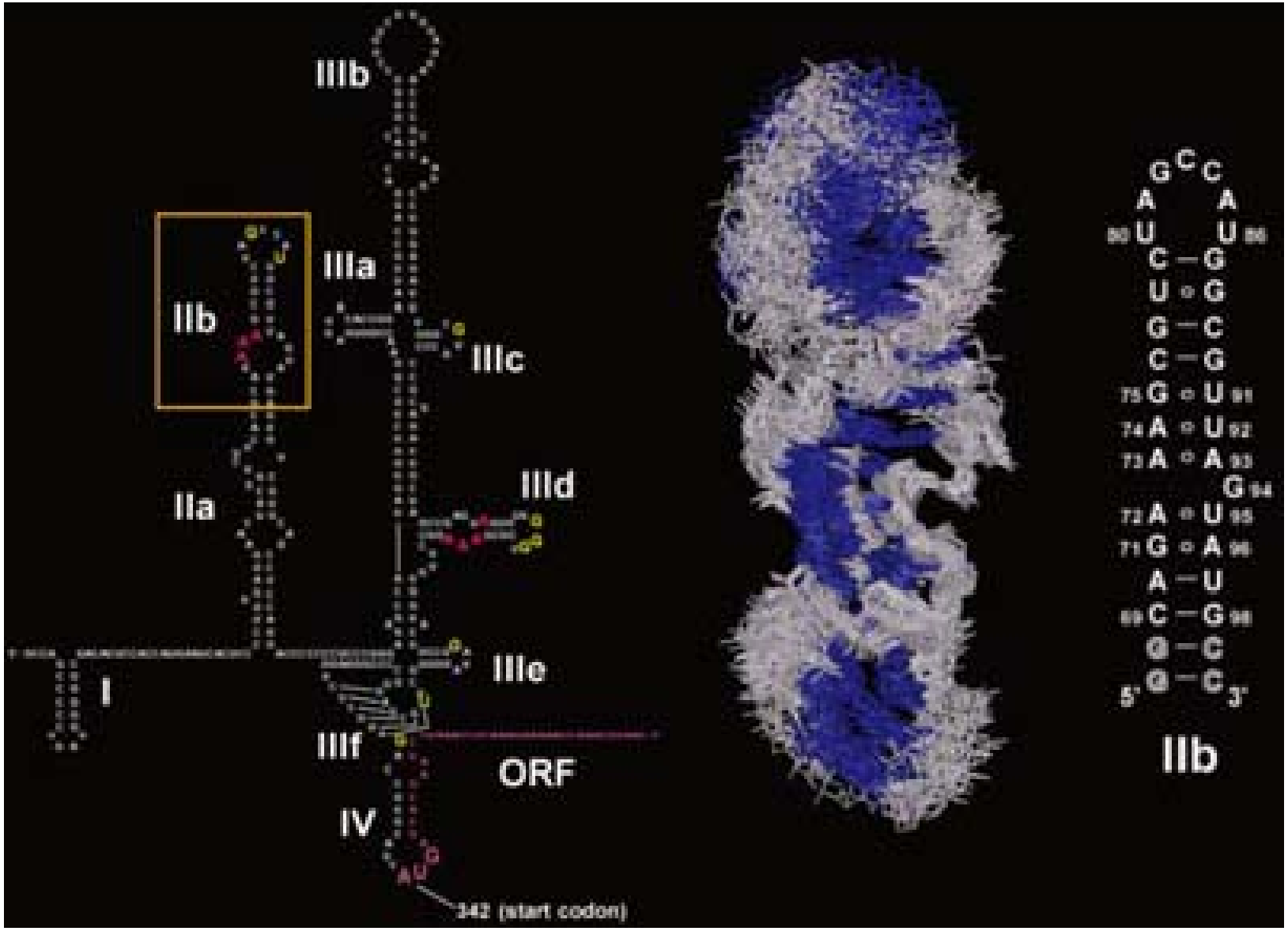
RNA Secondary Structures





E.coli 16S rRNA

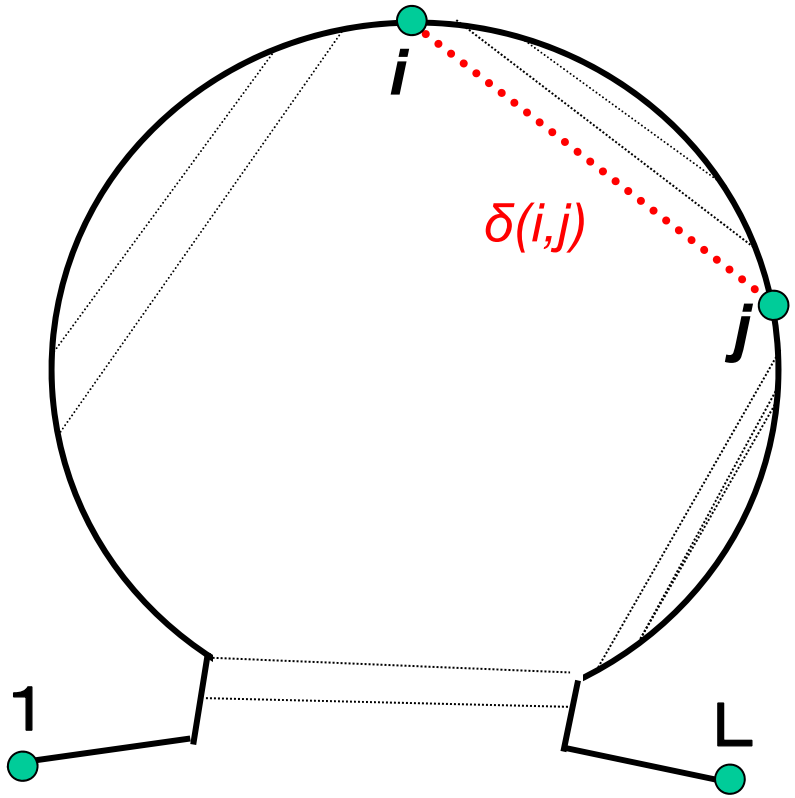
www.web.virginia.edu/Heidi/chapter12/chp12.htm



HCV virus RNA structure

http://puglisi.stanford.edu/research.html

Nussinov Algorithm
- Maximizing Base Pair Counts



maximize

$$E(1,L) = \sum_{1 \leqq i < j \leqq L} \delta(i, j)$$

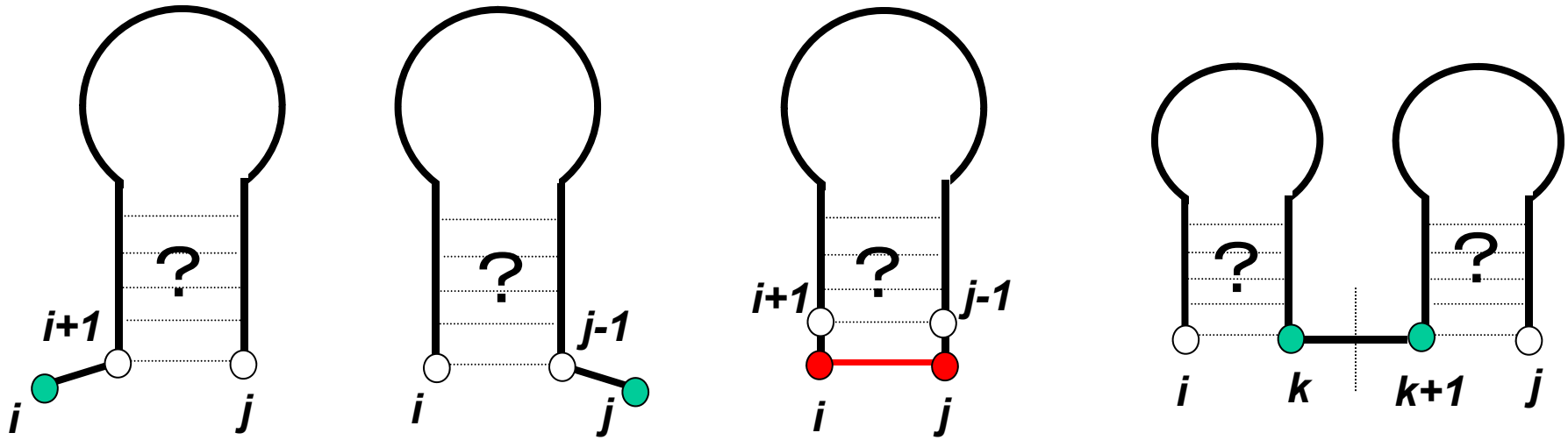
where $\delta(i,j)$ = 1 if base(i) and base(j) are hydrogen bonded, and $\delta(i,j)$ =0 otherwise.

Nussinov R, Piecznik G, Grigg JR and Kleitman DJ : "*Algorithms for loop matchings*". *SIAM Journal on Applied Mathematics* (1978).
Nussinov R and Jacobson AB : "Fast algorithm for predicting the secondary structure of single-stranded RNA", *Proc Natl Acad Sci USA*, 77(11):6309-13 (1980).

DP-based Algorithm by Nussinov

Recursive formulation of *E(i, j)*

(a) i unpaired **(b) j unpaired** **(c) i, j pair** **(d) bifurcation**



$$E(i,j) = \max \begin{cases} E(i+1,j), & \text{..(a)} \\ E(i,j-1), & \text{..(b)} \\ E(i+1,j-1) + \delta(i,j), & \text{..(c)} \\ \max_{i<k<j} \{ E(i,k) + E(k+1,j) \}. & \text{..(d)} \end{cases}$$

Initialization

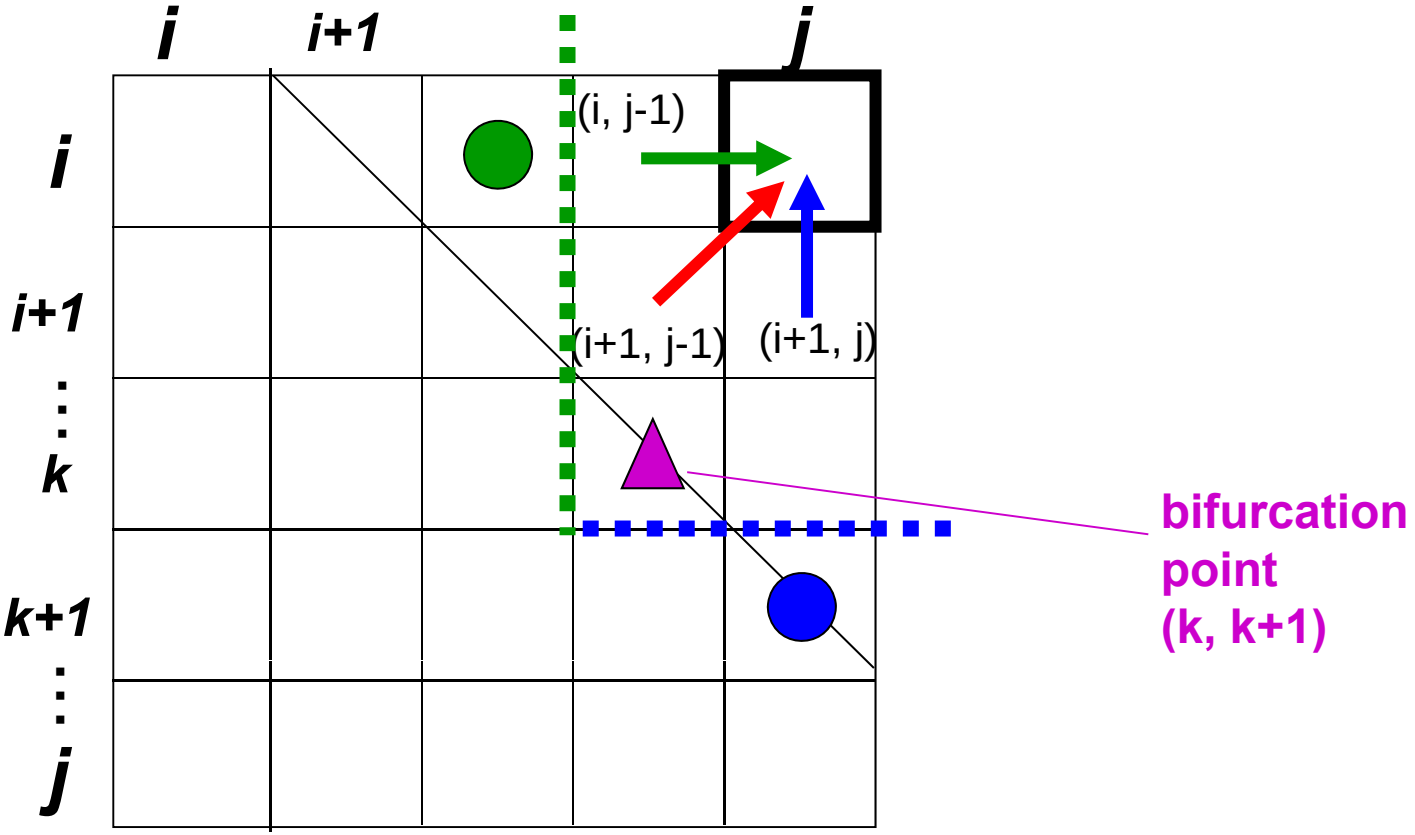
C	G	G	A	C	C	C	A	G	A	C	U	U	U	C
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

C	1
G	2
G	3
A	4
C	5
C	6
C	7
A	8
G	9
A	10
C	11
U	12
U	13
U	14
C	15

0														
0	0													
	0	0												
		0	0											
			0	0										
				0	0									
					0	0								
						0	0							
							0	0						
								0	0					
									0	0				
										0	0			
											0	0		
												0	0	
													0	0

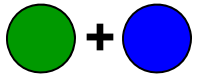
Matrix fill stage: diagonal to top-right

$$E(i, j) = \max \begin{cases} E(i+1, j) , & \text{..(a)} \\ E(i, j-1) , & \text{..(b)} \\ E(i+1, j-1) + \delta(i, j) , & \text{..(c)} \\ \max_{i<k<j} \{ E(i,k) + E(k+1, j) \} . & \text{..(d)} \end{cases}$$

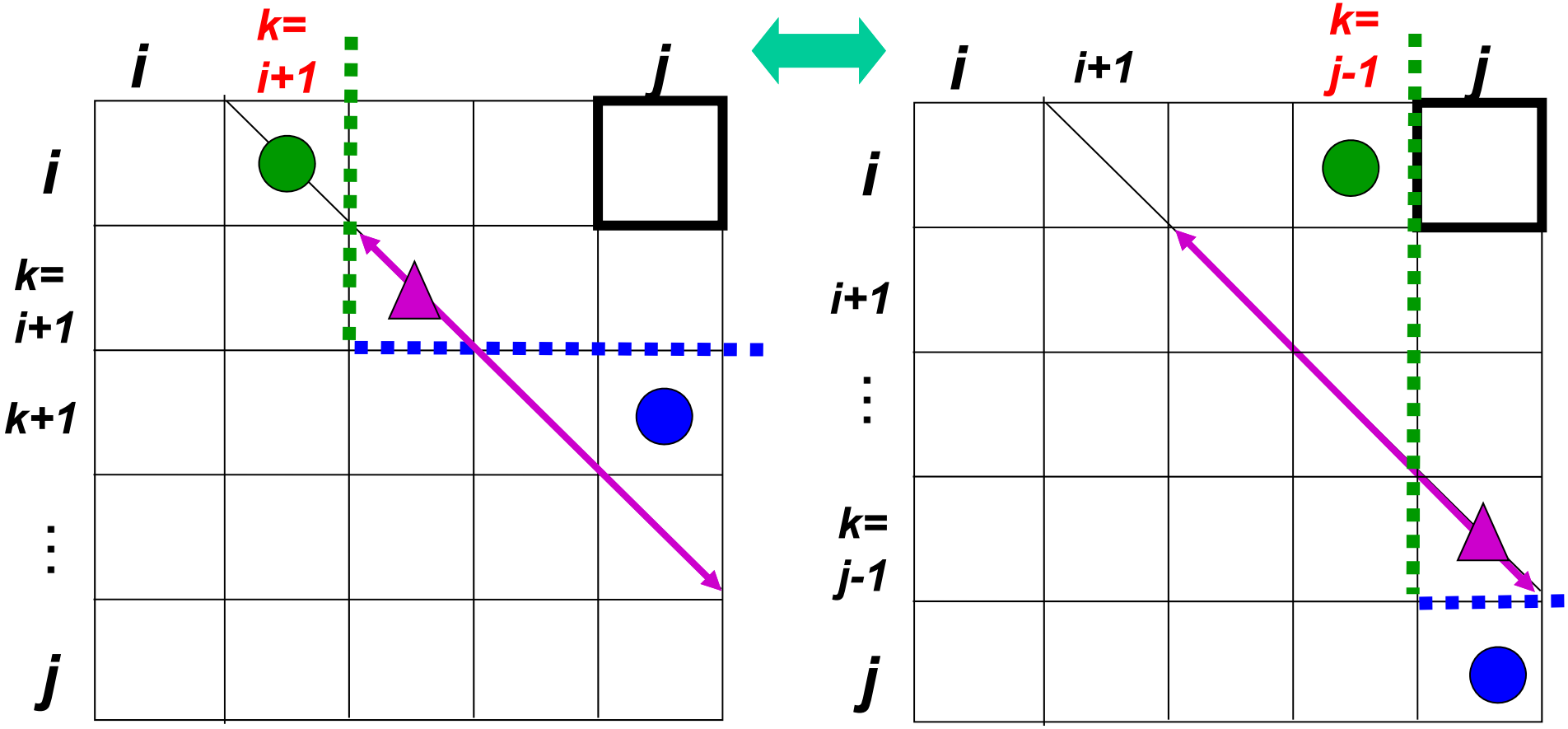


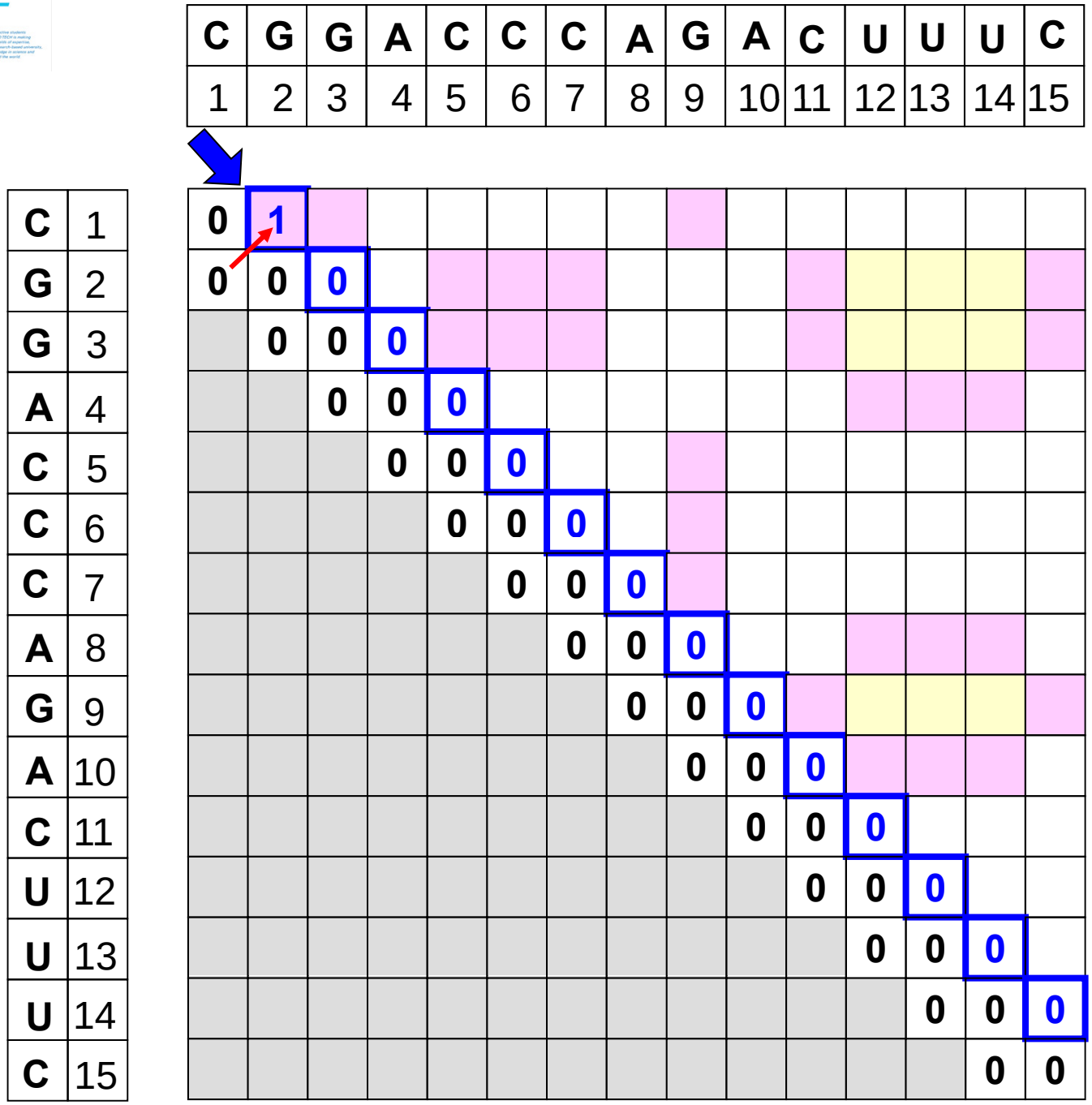
Matrix fill stage (cont'd)

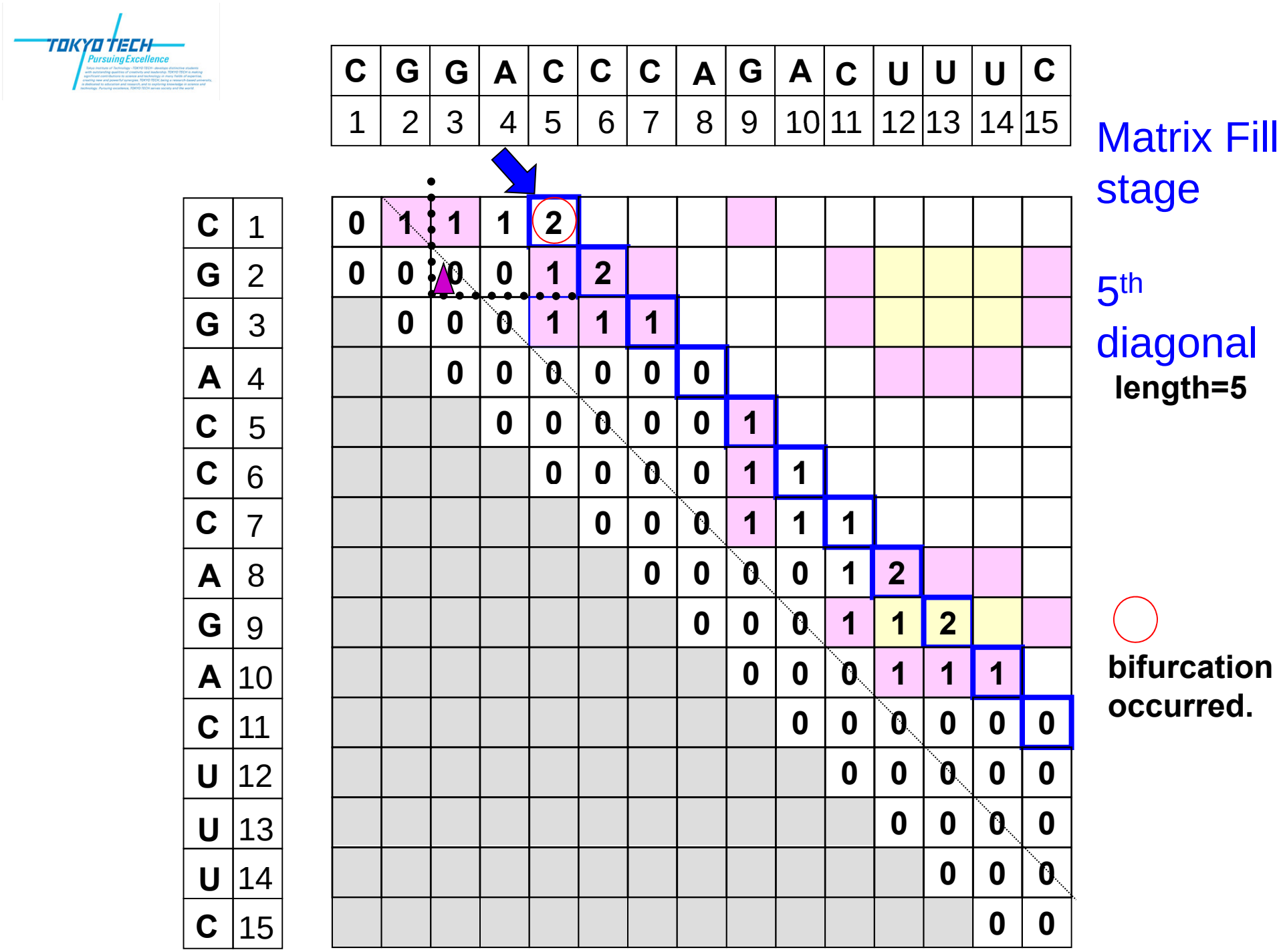
$$\max_{i<k<j} \{ E(i,k) + E(k+1, j) \} . \quad ..(d)$$

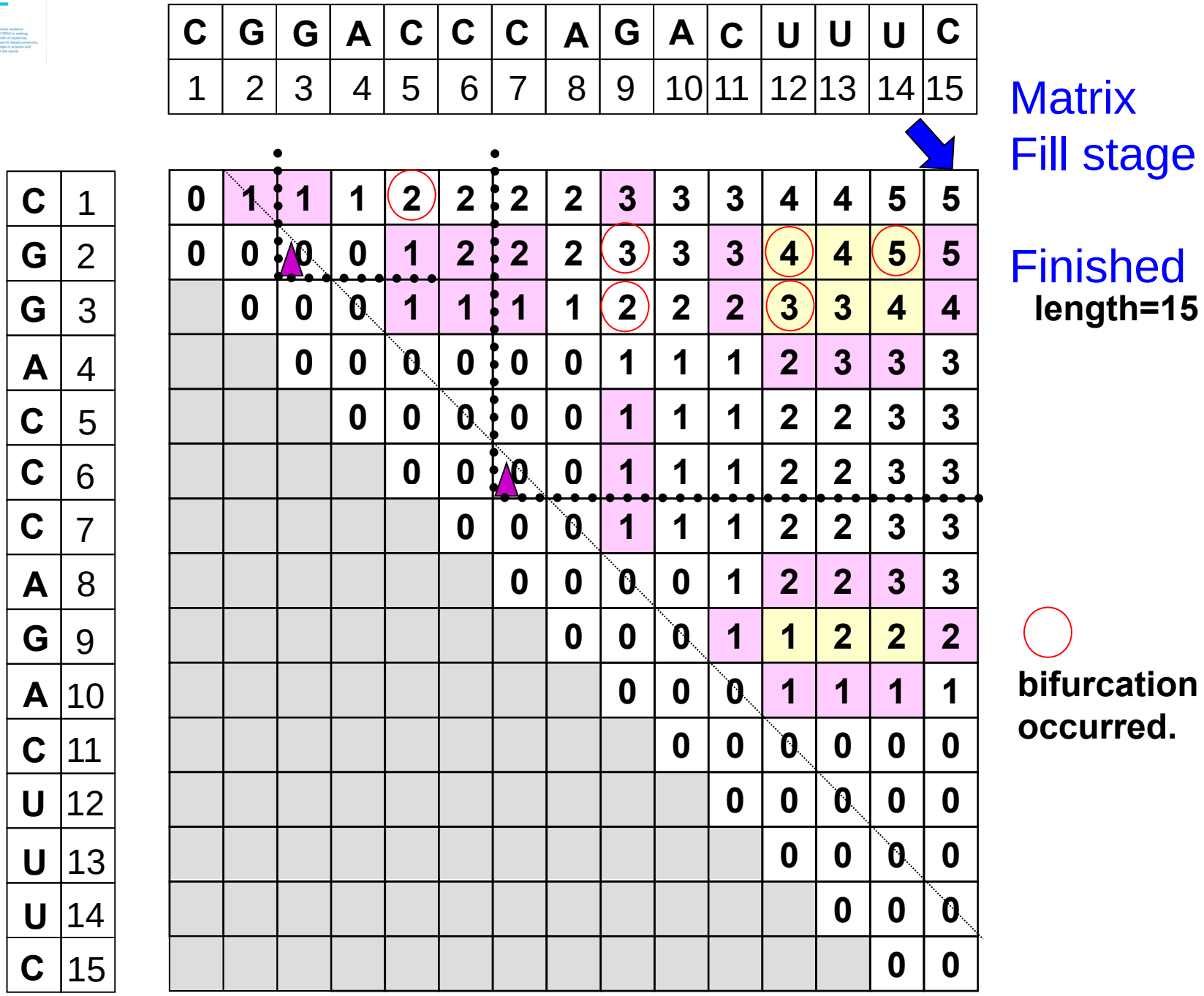


We need a scanning on the diagonal line for possible position of the bifurcation point ▲ . The coordinate *k* is within *i*<*k*<*j*.
The score is the maximum value of sum of score at ● and ● .




C G G A C C C A G A C U U U C
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
C 1
G 2
G 3
A 4
C 5
C 6
C 7
A 8
G 9
A 10
C 11
U 12
U 13
U 14
C 15
Matrix Fill stage
2nd diagonal
length=2


C G G A C C C A G A C U U U C
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
C 1
G 2
G 3
A 4
C 5
C 6
C 7
A 8
G 9
A 10
C 11
U 12
U 13
U 14
C 15
Matrix Fill stage
5th diagonal
length=5
bifurcation occurred.



Matrix Fill stage

Finished

length=15

○ **bifurcation occurred.**

Traceback stage

Initialization: **Push (1, L) onto stack**

Recursion: **Repeat until stack is empty:**

- **pop (i, j).**
- **if i >= j continue;**
  else if E(i+1, j) = E(i, j) push (i+1, j);
  else if E(i, j-1) = E(i, j) push (i, j-1);
  else if E(i+1, j-1)+δ(i,j) = E(i, j) :
    - **record i, j base pair;**
    - **push (i+1, j-1)**

  else for k=i+1 to j-1: if E(i,k)+E(k+1,j) = E(i,j):
    - **push (i, k);**
    - **push (k+1, j);**
    - **break;**

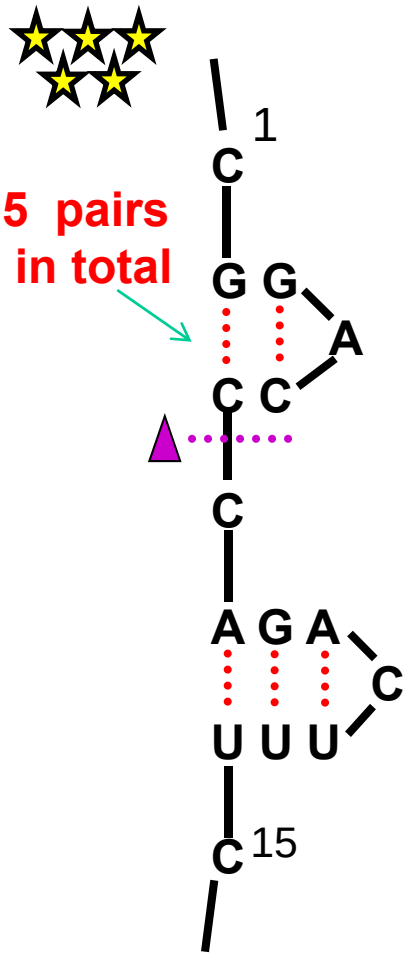
http://www.isrec.isb-sib.ch/topics_bioinfo_1/jan_12/RNA_folding_exercise.html

C	G	G	A	C	C	C	A	G	A	C	U	U	U	C
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

C	1
G	2
G	3
A	4
C	5
C	6
C	7
A	8
G	9
A	10
C	11
U	12
U	13
U	14
C	15

0	1	1	1	2	2	2	2	3	3	3	4	4	5	5
0	0	0	0	1	2	2	2	3	3	3	4	4	5	5
	0	0	0	1	1	1	1	2	2	2	3	3	4	4
		0	0	0	0	0	0	1	1	1	2	3	3	3
			0	0	0	0	0	1	1	1	2	2	3	3
				0	0	0	0	1	1	1	2	2	3	3
					0	0	0	1	1	1	2	2	3	3
						0	0	0	0	1	2	2	3	3
							0	0	0	1	1	2	2	2
								0	0	0	1	1	1	1
									0	0	0	0	0	0
										0	0	0	0	0
											0	0	0	0
												0	0	0
													0	0

Traceback stage



Some Extensions Required

1) different score for base pairs

G-C, C-G: +3
A-U, U-A: +2
G-U, U-G: +1
others: -1

2) minimum hairpin loop length

3) realistic energy function

→ Zuker (or Zuker-Stiegler) method

4) coping with “pseudoknot” structure

Zuker Method

- Energy score for “stacking region” (see adjacent two base pairs)

```
A    C    G    U
5' --> 3'
   AX
   UY
3' <-- 5'
0.4  0.4  0.4 -0.9
0.4  0.4 -2.1  0.4
0.4 -1.7  0.4 -0.5
-0.9 0.4 -1.0  0.4
```

...

```
A    C    G    U
5' --> 3'
   CX
   GY
3' <-- 5'
0.4  0.4  0.4 -1.8
0.4  0.4 -2.9  0.4
0.4 -2.0  0.4 -1.2
-1.7 0.4 -1.9  0.4
```

5’--A G--
--U C-- 5’
-1.7 kcal/mol

5’--A C--
--U G-- 5’
- 2.1 kcal/mol

Michael Zuker

MFOLD
prediction software

- Energy score for several loops

D.H. Mathews, J. Sabina, M. Zuker & D.H. Turner
Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure ***J. Mol. Biol.*** **288, 911-940 (1999)**