Sequence Alignment Algorithm

Bioinformatics Lecture

By Prof. Keun Woo Lee

경상대학교 생화학과 생물정보학연구실
Bioinformatics Lab http://bio.gsnu.ac.kr

Sequence Alignments: Terminology

용어정의:
-Similarity: 정량적, 통계적 유사성의 의미
-Homology: 정성적, 진화적거리 의미 (evolutional distance)

서열정렬의 종류:
-단순서열비교 (sequence comparison)
-동적프로그래밍(dynamic programming): LCS, Hidden Markov Chain
-포괄정렬 (global alignment): pairwise alignment, multiple alignment
-국부정렬 (local alignment): pairwise alignment, multiple alignment

Sequence Alignment Methods

Pairwise Alignment

Needleman, S. B.; Wunsch, C. D. "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins, " *J. Mol. Biol.*, **48**, 443-453 (1970).

Multiple Sequence Alignment

- Globally Optimal Multiple Alignment

Lipman, D.J.; Altschul, S.F.; Kececioglu, J.D. "A tool for multiple sequence alignment," *Proc. Natl. Acad. Sci. USA*, **86**, 4412-4415 (1989).

- Iterative Construction of a Multiple Alignment

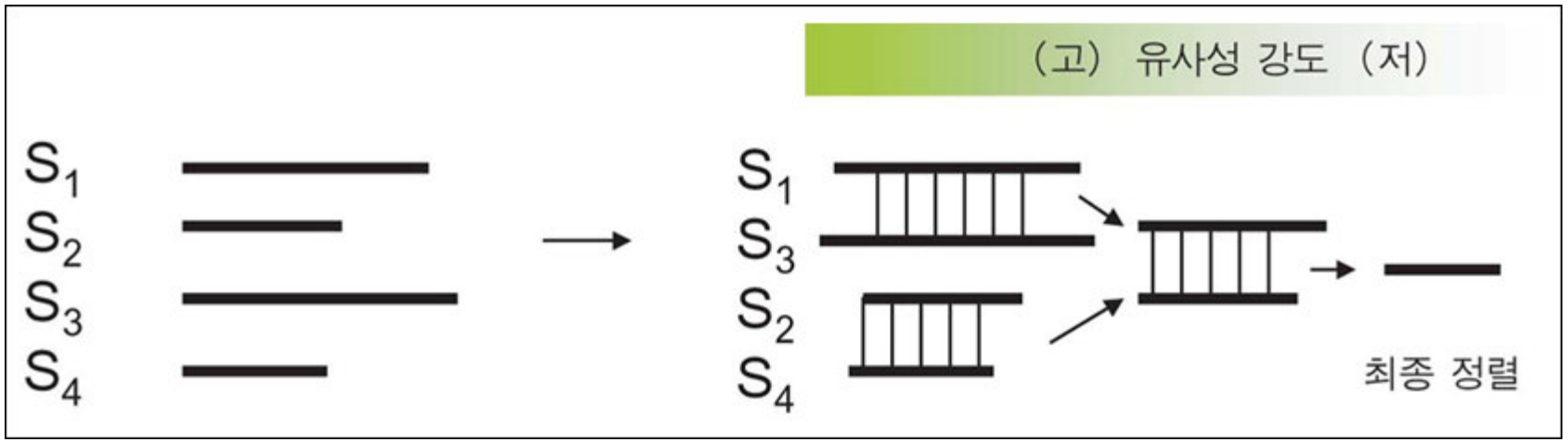
Berger, M.P.; Munson, P.J. "A novel randomized iterative strategy for aligning multiple protein sequences," *Comput Appl Biosci*, **7**, 479-484 (1991).

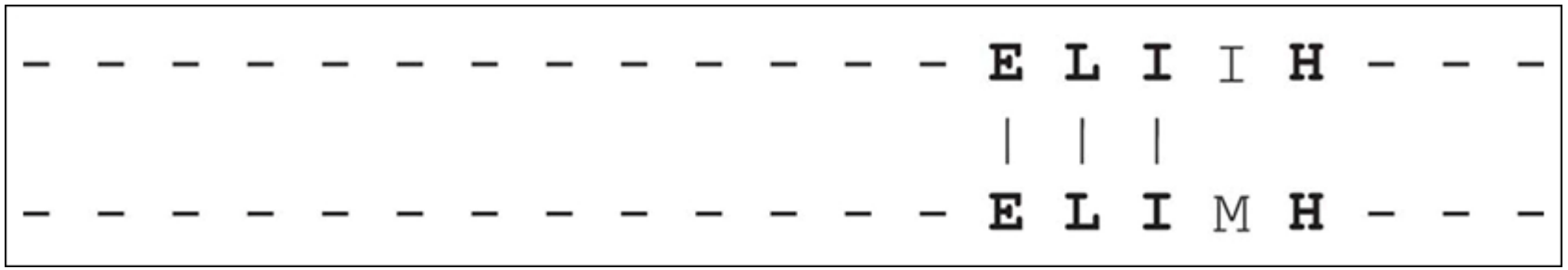
- Local Similarity Measures

Schuler, G.D.; Altschul, S.F.; Lipman, D.J. "A workbench for multiple alignment construction and analysis," *Proteins Struct. Func. Gen.*, **9**, 180-190 (1991).

Pairwise Alignment

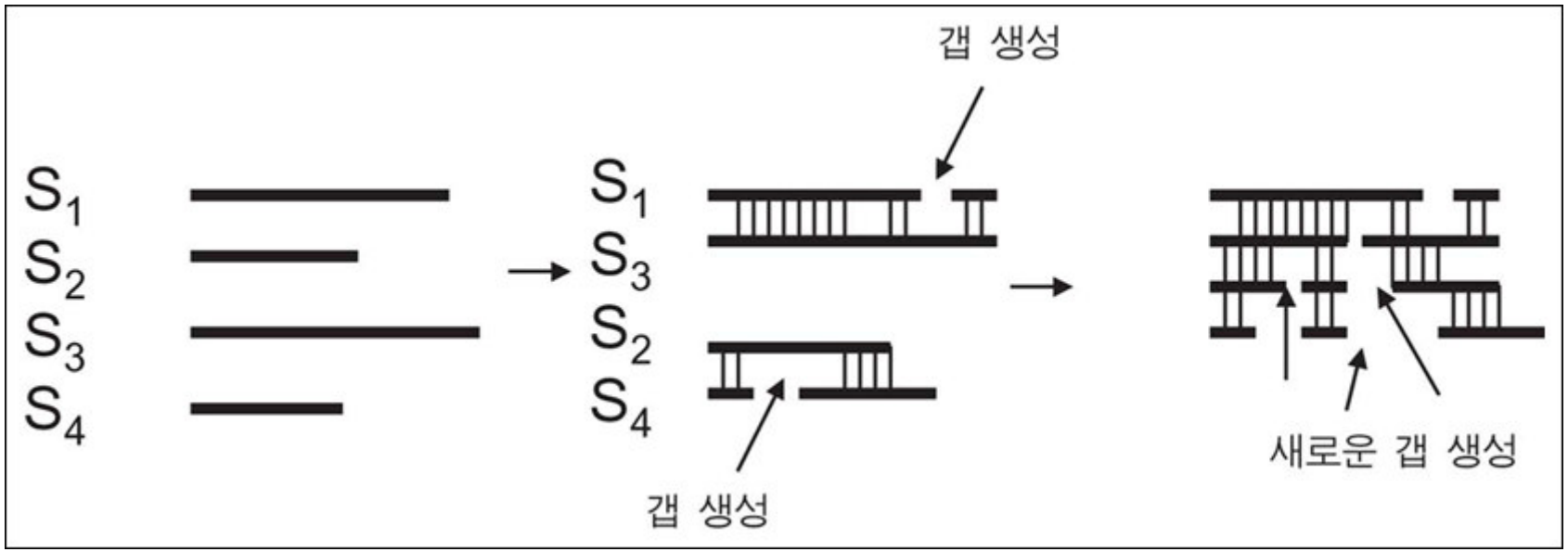
2개의 서열비교시





Multiple Alignment

3개 이상의 서열 동시 비교시



경상대학교 생화학과 생물정보학연구실
Bioinformatics Lab http://bio.gsnu.ac.kr

Needleman-Wunsch(NW) Algorithm

(ex)

A C F G S T V I Q N

and

C F G H A S T V Q N

Step 1. Set up Comparison Matrix Between the Two Sequences

	A	C	F	G	S	T	V	I	Q	N
C	0	1	0	0	0	0	0	0	0	0
F	0	0	1	0	0	0	0	0	0	0
G	0	0	0	1	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0
A	1	0	0	0	0	0	0	0	0	0
S	0	0	0	0	1	0	0	0	0	0
T	0	0	0	0	0	1	0	0	0	0
V	0	0	0	0	0	0	1	0	0	0
Q	0	0	0	0	0	0	0	0	1	0
N	0	0	0	0	0	0	0	0	0	1

Step 2. Processing the Comparison Matrix

	A	C	F	G	S	T	V	I	Q	N
C										
F										
G										
H										
A										
S					5					
T						4	2	2	1	0
V						2	3	2	1	0
Q						1	1	1	2	0
N						0	0	0	0	1

Rule:

$M_{i,j} = MAX[(M_{i-1,j-1} + S_{ij}), (M_{i,j-1} + w), (M_{i-1,j-1} + w)]$

S_{ij}=1 for match 0 for mismatch (친화도 affinity),
w=gap in the sequence

Processed Comparison Matrix

	A	C	F	G	S	T	V	I	Q	N
C	7	8	6	5	4	3	2	2	1	0
F	6	6	7	5	4	3	2	2	1	0
G	5	5	5	6	4	3	2	2	1	0
H	5	5	5	5	4	3	2	2	1	0
A	6	5	5	5	4	3	2	2	1	0
S	4	4	4	4	5	3	2	2	1	0
T	3	3	3	3	3	4	2	2	1	0
V	2	2	2	2	2	2	3	2	1	0
Q	1	1	1	1	1	1	1	1	2	0
N	0	0	0	0	0	0	0	0	0	1

Step 3. Finding Maximum Pathway

	A	C	F	G	S	T	V	I	Q	N
C	7	8	6	5	4	3	2	2	1	0
F	6	6	7	5	4	3	2	2	1	0
G	5	5	5	6	4	3	2	2	1	0
H	5	5	5	5	4	3	2	2	1	0
A	6	5	5	5	4	3	2	2	1	0
S	4	4	4	4	5	3	2	2	1	0
T	3	3	3	3	3	4	2	2	1	0
V	2	2	2	2	2	2	3	2	1	0
Q	1	1	1	1	1	1	1	1	2	0
N	0	0	0	0	0	0	0	0	0	1

Step 4. Final Maximum Pathway and Corresponding Sequence Alignment

(a)

	A	C	F	G	S	T	V	I	Q	N
C	7	**8**	6	5	4	3	2	2	1	0
F	6	6	**7**	5	4	3	2	2	1	0
G	5	5	5	**6**	4	3	2	2	1	0
H	5	5	5	5	4	3	2	2	1	0
A	6	5	5	5	4	3	2	2	1	0
S	4	4	4	4	**5**	3	2	2	1	0
T	3	3	3	3	3	**4**	2	2	1	0
V	2	2	2	2	2	2	**3**	2	1	0
Q	1	1	1	1	1	1	1	1	**2**	0
N	0	0	0	0	0	0	0	0	0	**1**

(b)

A	C	F	G	—	—	S	T	V	I	Q	N
	C	F	G	H	A	S	T	V	—	Q	N

Final result

Needleman-Wunsch Algorithm: 예시2

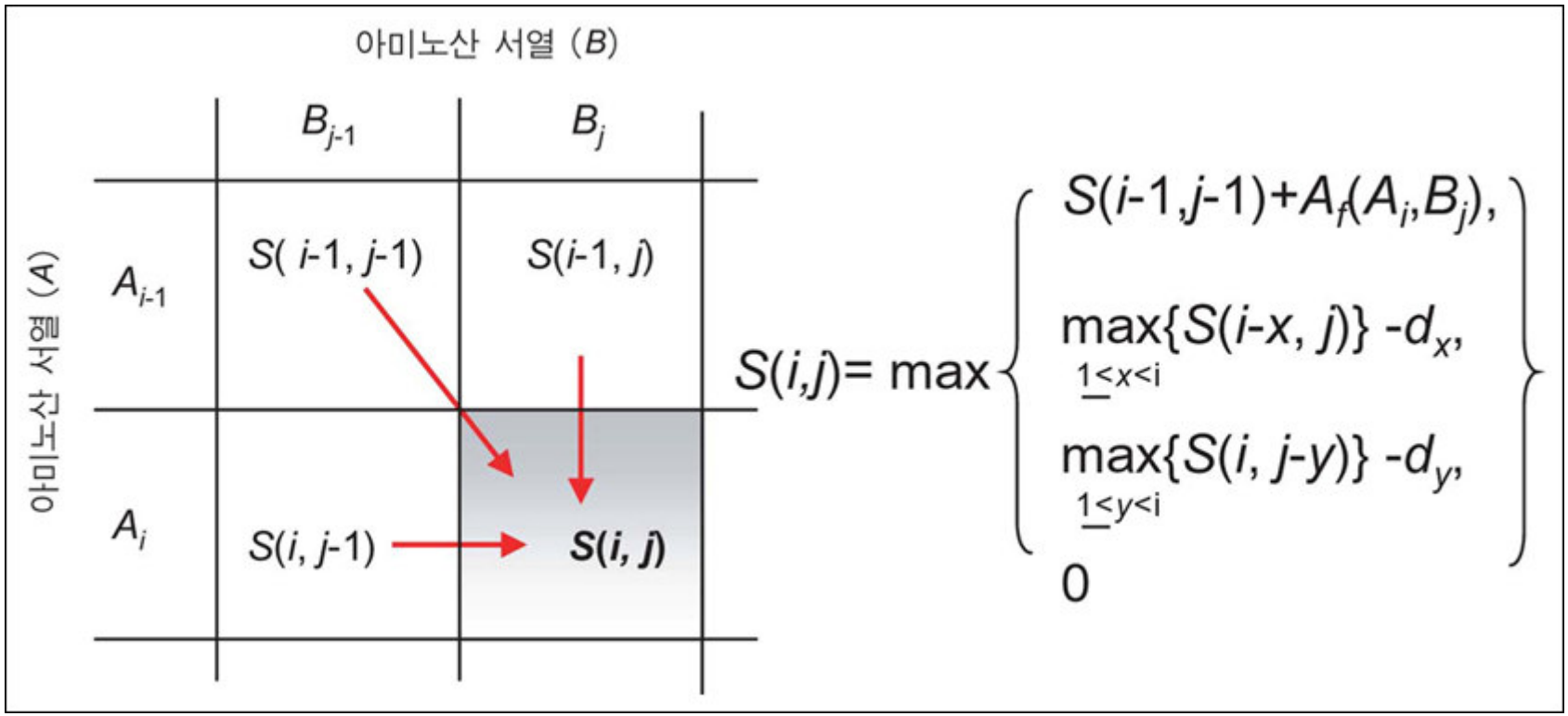
(1) NW(7,8)의 행렬값 연산 예

	W	S	R	H	L	H	Q	R	Y	P	H	D	A
S	0	1	0	0	0	0	0	0	0	0	0	0	0
D	0	0	1	1	1	1	1	1	1	1	1	2	1
R	0	0	2	1	1	1	1	2	1	1	1	1	2
H	0	0	1	3	2	3	2	2	2	2	3	2	2
K	0	0	1	2	3	3	3	3	3	3	3	3	3
H	0	0	1	3	3	4	3	3	3	3	4	3	3
R	0	0	2	2	3	3	4	?					
P													
Y													
H													
Y													
A													

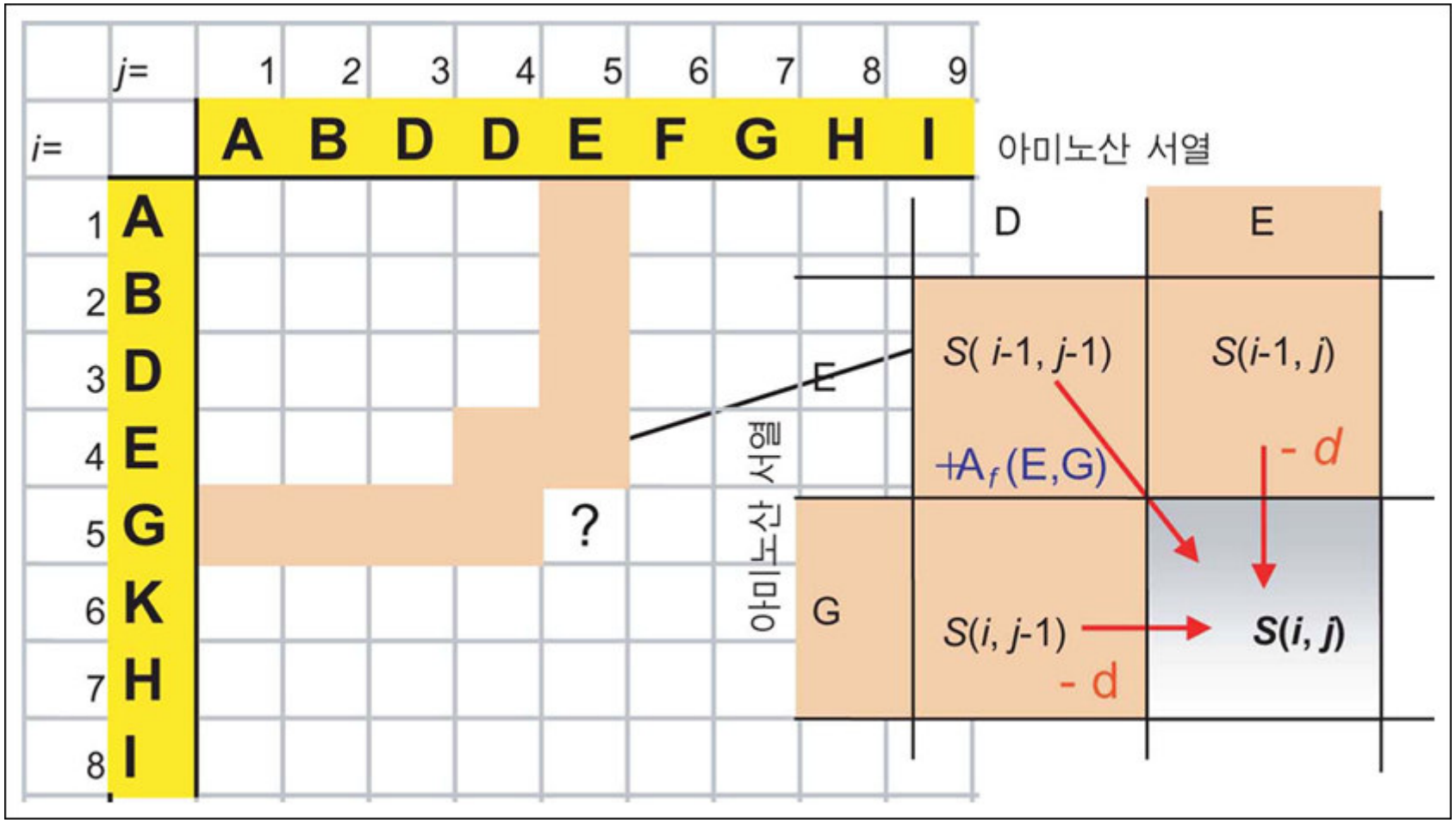
(2) 전체 행렬값 결과 및 탐색

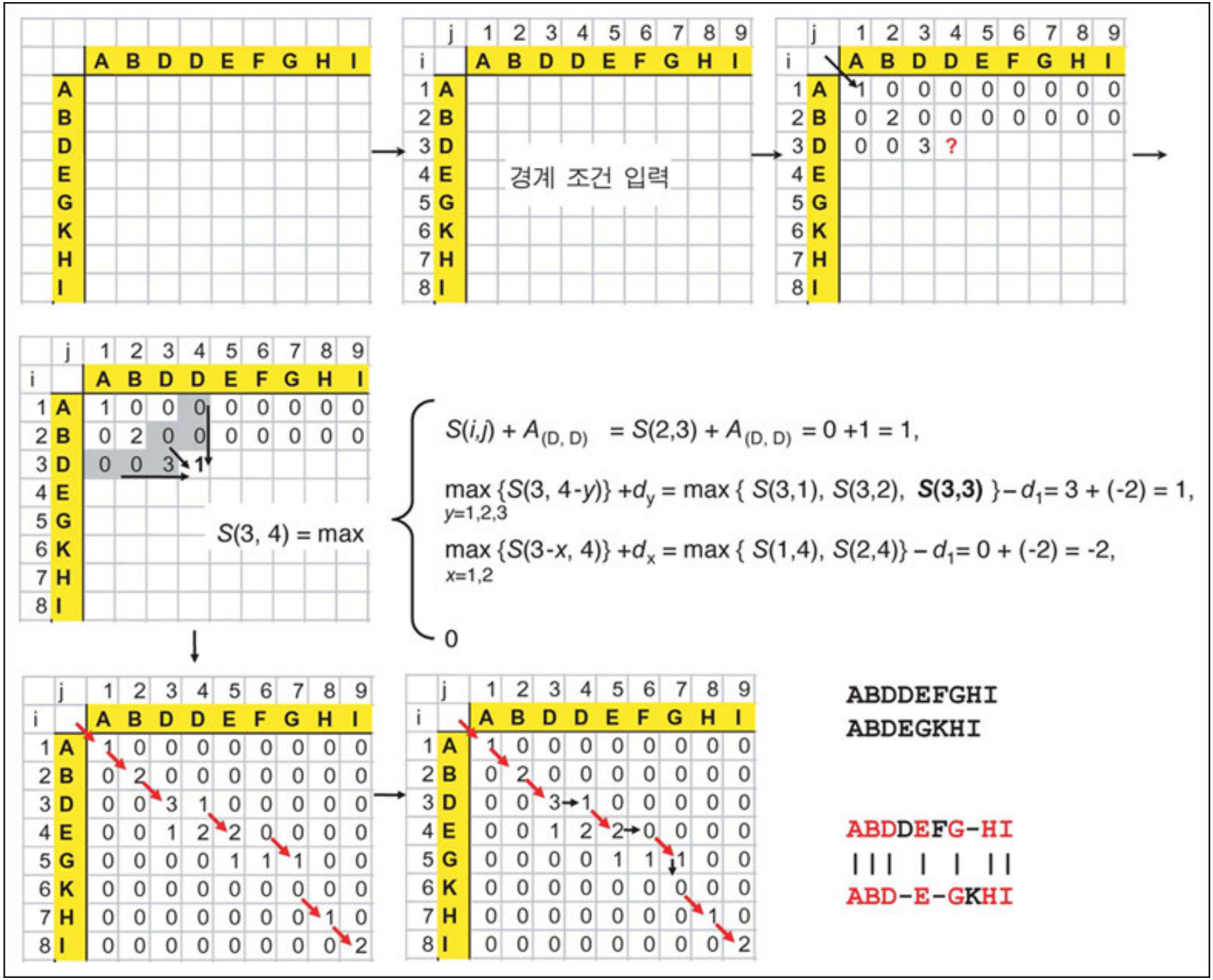
	W	S	R	H	L	H	Q	R	Y	P	H	D	A
S	0	1	0	0	0	0	0	0	0	0	0	0	0
D	0	0	1	1	1	1	1	1	1	1	1	2	1
R	0	0	2	1	1	1	1	2	1	1	1	1	2
H	0	0	1	3	2	3	2	2	2	2	3	2	2
K	0	0	1	2	3	3	3	3	3	3	3	3	3
H	0	0	1	3	3	4	3	3	3	3	4	3	3
R	0	0	2	2	3	3	4	5	4	4	4	4	4
P	0	0	1	2	3	3	4	4	5	*6	5	5	5
Y	0	0	1	2	3	3	4	4	6	5	6	6	6
H	0	0	1	3	3	4	4	4	5	6	7	6	6
Y	0	0	1	2	3	3	4	4	6	6	6	7	7
A	0	0	1	2	3	3	4	4	5	6	6	7	8

Smith-Waterman(SW) Algorithm

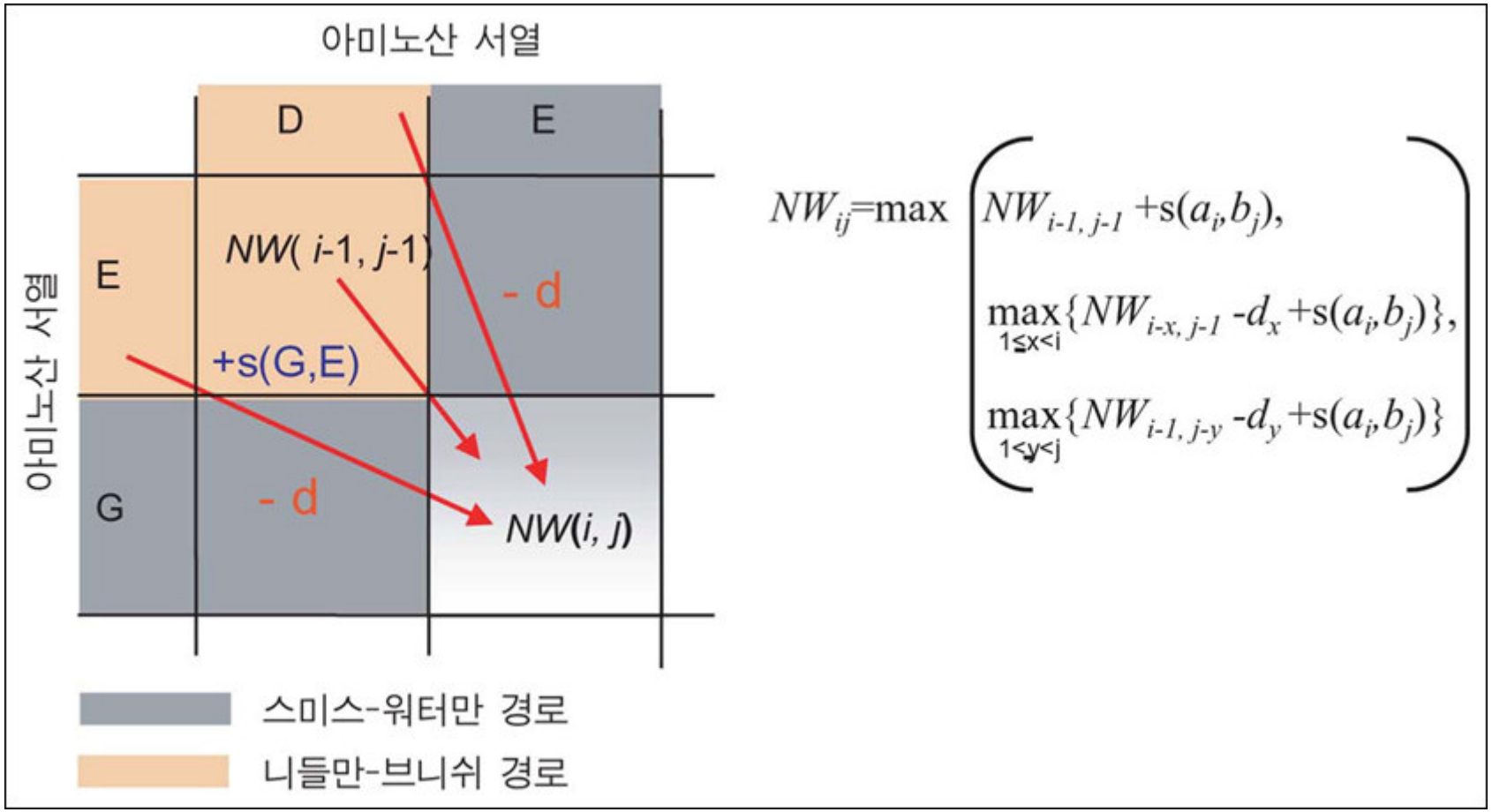


Smith-Waterman Algorithm




A B D D E F G H I
A B D E G K H I
j 1 2 3 4 5 6 7 8 9
i
경계 조건 입력
?
S(3, 4) = max
S(i,j) + A(D, D) = S(2,3) + A(D, D) = 0 +1 = 1,
max {S(3, 4-y)} +dy = max { S(3,1), S(3,2), S(3,3) } – d1= 3 + (-2) = 1,
y=1,2,3
max {S(3-x, 4)} +dx = max { S(1,4), S(2,4)} – d1= 0 + (-2) = -2,
x=1,2
0
ABDDEFGHI
ABDEGKHI
ABDDEFG-HI
ABD-E-GKHI

NW vs. SW Algorithm



$$NW_{ij}=\max\left(\begin{array}{l} NW_{i-1,\,j-1}+\mathrm{s}(a_i,b_j), \\ \max\limits_{1\le x<i}\{NW_{i-x,\,j-1}-d_x+\mathrm{s}(a_i,b_j)\}, \\ \max\limits_{1<y<j}\{NW_{i-1,\,j-y}-d_y+\mathrm{s}(a_i,b_j)\} \end{array}\right)$$