- 編集距離の計算にトレースバックを追加すれば配列の整列（アラインメント）が計算できる．
- トレースバックの方法は、次ページ以降のNeedleman-Wunsch法を参考にすること．（距離の計算方法が少し違うのみで、基本的なアルゴリズムの構造は、編集距離の計算と同じ．）

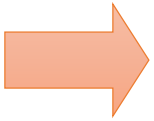
		T	A	T	A	T	A	T	A
	0	1	2	3	4	5	6	7	8
A	1	1	1	2	3	4	5	6	7
T	2	1	2	1	2	3	4	5	6
G	3	2	2	2	2				
T	4	3	3						
A									
T									
A									
T									

編集距離の計算では、距離のみを記録していたが、どこから来たかも記録する．この場合は、斜め上から来たことを記録しておく．（つまり，AとGが並ぶことを記録）

L(3,4)=2
L(2,3)+M(G,A): 2
L(2,4)+1: 3
L(3,3)+1: 3

配列の整列（アラインメント）

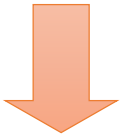
HEAGAWGHEE
PAWHEAE



HEAGAWGHE-E
--P-AW-HEAE

SMEDKSNVKAIWGKASGHLEEYGAEALERMFCAYPQTKIYFPHFDM

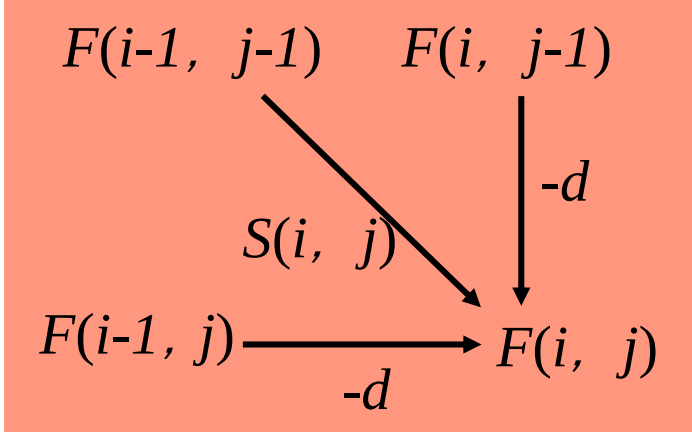
SPADKTNVKDKIGGHAGALERTFASFPTTKTYFPHFDL



SMEDKSNVKAIWGKASGHLEEYGAEALERMFCAYPQTKIYFPHFDM

SPADKTNVK---DKIGGHAG-----ALERTFASFPTTKTYFPHFDL

Needleman-Wunsch



$$F(0, \quad j) = j \times (-d)$$
$$F(i, \quad 0) = i \times (-d)$$

$$F(i, \quad j) = \max \begin{cases} F(i\text{-}1, \quad j\text{-}1) + S(x_i, \quad y_j) \\ F(j\text{-}1, \quad j) - d \\ F(i, \quad j\text{-}1) - d \end{cases}$$

Gap penalty

$$P(i, \quad j) = \begin{cases} diag : if \ F(i\text{-}1, \quad j\text{-}1) + S(x_i, \quad y_j) \ is \ max \\ up \quad : if \ F(j\text{-}1, \quad j) - d \ is \ max \\ left \ : if \ F(i, \quad j\text{-}1) - d \ is \ max \end{cases}$$

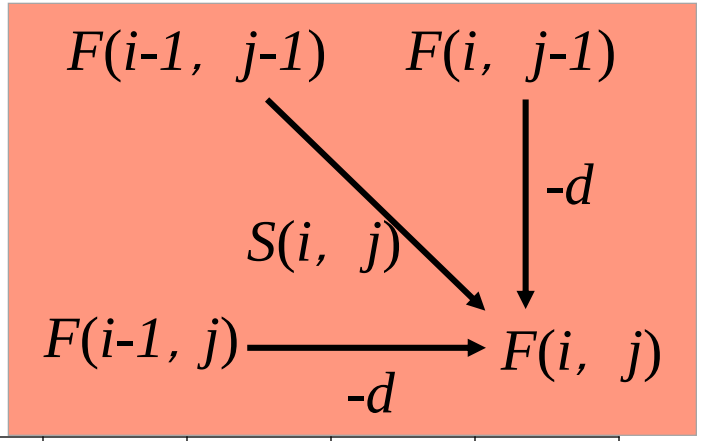
Storing pointer for traceback.

関数Sの例： 例えば 文字AとAをマッチさせる場合のスコアS(A, A)は4

	A	G	S	T	N	D	E	Q	K	R	H	M	I	L	V	F	Y	W	P	C
A	4	0	1	0	-2	-2	-1	-1	-1	-1	-2	-1	-1	-1	0	-2	-2	-3	-1	0
G	0	6	0	-2	0	-1	-2	-2	-2	-2	-2	-3	-4	-4	-3	-3	-3	-2	-2	-3
S	1	0	4	1	1	0	0	0	0	-1	-1	-1	-2	-2	-2	-2	-2	-3	-1	-1
T	0	-2	1	5	0	-1	-1	-1	-1	-1	-2	-1	-1	-1	0	-2	-2	-2	-1	-1
N	-2	0	1	0	6	1	0	0	0	0	1	-2	-3	-3	-3	-3	-2	-4	-2	-3
D	-2	-1	0	-1	1	6	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4	-1	-3
E	-1	-2	0	-1	0	2	5	2	1	0	0	-2	-3	-3	-2	-3	-2	-3	-1	-4
Q	-1	-2	0	-1	0	0	2	5	1	1	0	0	-3	-2	-2	-3	-1	-2	-1	-3
K	-1	-2	0	-1	0	-1	1	1	5	2	-1	-1	-3	-2	-2	-3	-2	-3	-1	-3
R	-1	-2	-1	-1	0	-2	0	1	2	5	0	-1	-3	-2	-3	-3	-2	-3	-2	-3
H	-2	-2	-1	-2	1	-1	0	0	-1	0	8	-2	-3	-3	-3	-1	2	-2	-2	-3
M	-1	-3	-1	-1	-2	-3	-2	0	-1	-1	-2	5	1	2	1	0	-1	-1	-2	-1
I	-1	-4	-2	-1	-3	-3	-3	-3	-3	-3	-3	1	4	2	3	0	-1	-3	-3	-1
L	-1	-4	-2	-1	-3	-4	-3	-2	-2	-2	-3	2	2	4	1	0	-1	-2	-3	-1
V	0	-3	-2	0	-3	-3	-2	-2	-2	-3	-3	1	3	1	4	-1	-1	-3	-2	-1
F	-2	-3	-2	-2	-3	-3	-3	-3	-3	-3	-1	0	0	0	-1	6	3	1	-4	-2
Y	-2	-3	-2	-2	-2	-3	-2	-1	-2	-2	2	-1	-1	-1	-1	3	7	2	-3	-2
W	-3	-2	-3	-2	-4	-4	-3	-2	-3	-3	-2	-1	-3	-2	-3	1	2	11	-4	-2
P	-1	-2	-1	-1	-2	-1	-1	-1	-1	-2	-2	-2	-3	-3	-2	-4	-3	-4	7	-3
C	0	-3	-1	-1	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2	-3	9

DP行列の計算

Gapのペナルティは -6 とする.



MAX

D: -5+(-2) = -7
U: -11-6 = -17
L: -8 - 6 = -14

					G	A	W	G	H	E	E
	0 ←	−6 ←		−18 ←	−24 ←	−30 ←	−36 ←	−42 ←	−48 ←	−54 ←	−60
P	↑ −6	↖ −2	↖ −7		−19 ←	↖ −25	← −31	← −37	← −43	↖ −49	↖ −55
A	↑ −12	↖ −8	↖ −3		−9	↖ −15	← −21	← −27	← −33	← −39	← −45
W	↑ −18	↖ −14	↑ −9	↖ −6	↖-5	←-11					
H	↑ −24	↖ −10	↖ −14	↖ −11	↖-8	↖					
E											
A											
E											

トレースバック

2つの配列のアラインメントは，右下のセルからポインタをたどっていけば計算できる．

HEAGAWGHE−E

−−P−AW−HEAE

		H	E	A	G	A	W	G	H	E	E
	0 ←	−6 ←	−12 ←	−18 ←	−24 ←	−30 ←	−36 ←	−42 ←	−48 ←	−54 ←	−60
P	↑ −6	↖ −2	↖ −7	↖ −13	← −19	↖ −25	← −31	← −37	← −43	↖ −49	↖ −55
A	↑ −12	↖ −8	↖ −3	↖ −3	← −9	↖ −15	← −21	← −27	← −33	← −39	← −45
W	↑ −18	↖ −14	↑ −9	↖ −6	↖ −5	← −11	↖ −4	← −10	← −16	← −22	← −28
H	↑ −24	↖ −10	↖ −14	↖ −11	↖ −8	↖ −7	↑ −10	↖ −6	↖ −2	← −8	← −14
E	↑ −30	↑ −16	↖ −5	← −11	↖ −13	↖ −9	↖ −10	↖ −12	↖ −6	↖ 3	↖ −3
A	↑ −36	↑ −22	↑ −11	↖ −1	← −7	↖ −9	↖ −12	↖ −10	↑ −12	↑ −3	↖ 2
E	↑ −42	↑ −28	↖ −17	↑ −7	↖ −3	↖ −8	↖ −12	↖ −14	↖ −10	↖ −7	↖ 2

- 今回の課題では，挿入・削除は高々5%程度.
- しかし，とても長い配列と短い配列を整列させるのに編集距離と同じように配列間の距離を計算してしまうと，良いアラインメント（挿入・削除がバラバラに入る）と悪いアラインメント（短い配列の中に挿入・削除があまり入らない）の区別がつかない.

- 例）(1)と(2) の編集距離は同じだが，今回は(2)のアラインメントを求めたい.

(1)
```
b - b - - - a - - - a - b c b
b c b c b b a b c b a a b b b
```

(2)
```
- - - - b b a a b c b - - - -
b c b c b b a - b c b a a b b
```

- これは，以下のように計算を工夫することで解決できる．
  - DP行列を計算するときに1行目と1列目の値をすべて0にする．（長い配列の途中から整列を開始しても距離が大きくならないようにする．）
  - 短い配列の長さをLとしたときに，一番右の列のL+1行目以降で最も小さい値を示す要素からトレースバックする．

- 次ページに計算の例を示したので，確認してみてください．

		b	b	a	a	b	c	b
	0	0	0	0	0	0	0	0
b	0	0	0	1	1	0	1	0
c	0	1	1	1	2	1	0	1
b	0	0	1	2	2	2	1	0
c	0	1	1	2	3	3	2	1
b	0	0	1	2	3	3	3	2
b	0	0	0	1	2	3	4	3
a	0	1	1	0	1	2	3	4
b	0	0	1	1	1	1	2	3
c	0	1	1	2	2	2	1	2
b	0	0	1	2	3	2	2	1
a	0	1	1	1	2	3	3	2
a	0	1	2	1	1	2	3	3
b	0	0	1	2	2	2	2	3
b	0	0	0	1	2	2	2	2
a	0	1	1	2	2	3	3	3

		b	b	a	a	b	c	b
		–	–	–	–	–	–	–
b	|	*	*	–	|	*	–	*
c	|	–	|	*	–	|	*	–
b	|	*	–	–	*	|	|	*
c	|	–	*	–	–	|	|	|
b	|	*	–	–	–	*	|	|
b	|	*	*	–	–	–	–	|
a	|	–	|	*	–	–	–	–
b	|	*	–	|	*	*	–	–
c	|	–	*	–	|	|	*	–
b	|	*	–	–	–	*	|	*
a	|	–	*	*	–	–	|	|
b	|	*	–	–	|	*	–	–
b	|	*	*	–	–	|	*	*
a	|	–	|	*	–	–	–	|

8行目以降で最小の値なので，
この位置からトレースバック．

（'*'は斜め上，'|'は上，'–'は左を示す．）

b b a a b c b

b c b c b b a – b c b a a b b