

4. Use the dataset in Exercise 3. Let the similarity(P4,P6)=similarity(P5,P7)=0.9, and for any other page pairs, similarity(Pi,Pj)=1 if i=j, and similarity(Pi,Pj)=0 if i!=j. Use the Needleman-Wunsch Algorithm ([LuDM05]) to determine the best alignment of S3 and S4.

Data Set:

S3: <P1,P6,P1,P4>

S4: <P5,P4,P1,P6,P7>

Needleman-Wunsch Alignment Algorithm

Consider two sequences $X_1 \dots X_i$ and $Y_1 \dots Y_j$, the optimal alignment score $A(i, j)$ is the maximum of the following

- $A(i-1, j-1) + s(X_i, Y_j)$
- $A(i-1, j) + d$
- $A(i, j-1) + d$

$s(X_i, Y_j)$ is the similarity between X_i and Y_j , and d is the score of aligning X_i or Y_j with a gap.

Compute alignment score

		Y_1	...	Y_{j-1}	Y_j	...	Y_n
	0	-d	...	$-(j-1)d$	-jd	...	-nd
X_1	-d						
...	...						
X_{i-1}	$-(i-1)d$			$A(i-1, j-1)$	$A(i-1, j)$		
X_i	-id			$A(i, j-1)$	$A(i, j)$		
...	...						
X_m	-md						$A(m, n)$

Fig 3. Computing optimal alignment of two sequences using Needleman-Wunsch algorithm

Similarity(P4,P6) = Similarity(P5,P7) = 0.9

Similarity(Pi, Pj) = 0 if i!= j

Similarity(Pi, Pj) =1 if i=j

Assume $d = -10$

Pair	$A(i-1, j-1) + S(X_i, Y_j)$	$A(i-1, j) + d$	$A(i, j-1) + d$	Alignment Score
A(P1,P5)	$0+0=0$	$10+(-10)=0$	$10+(-10)=0$	0
A(P1,P4)	$10+0=10$	$0+(-10)=-10$	$0+(-10)=-10$	10
A(P6,P5)	$10+0=10$	$20+(-10)=10$	$0+(-10)=-10$	10
A(P6,P4)	$0+0=0$	$10+(-10)=0$	$10+(-10)=0$	0
A(P1,P1)	$20+1=21$	$10+(-10)=0$	$30+(-10)=20$	21
A(P1,P5)	$20+0=20$	$30+(-10)=20$	$10+(-10)=0$	20
A(P4,P5)	$30+0=30$	$40+(-10)=30$	$20+(-10)=10$	30
A(P1,P6)	$30+0=30$	$21+(-10)=11$	$40+(-10)=30$	30
A(P6,P3)	$10+0=10$	$0+(-10)=-10$	$21+(-10)=11$	11
A(P1,P4)	$10+0=10$	$20+(-10)=10$	$0+(-10)=-10$	10
A(P4,P4)	$20+1=21$	$30+(-10)=20$	$10+(-10)=0$	21
A(P1,P7)	$40+0=40$	$30+(-10)=20$	$50+(-10)=40$	40
A(P6,P6)	$21+1=22$	$11+(-10)=1$	$30+(-10)=20$	22
A(P1,P1)	$0+1=1$	$10+(-10)=0$	$11+(-10)=1$	11
A(P6,P5)	$30+0=30$	$22+(-10)=12$	$40+(-10)=30$	30
A(P1,P6)	$11+0=11$	$1+(-10)=-9$	$22+(-10)=12$	12
A(P4,P1)	$10+0=10$	$21+(-10)=11$	$1+(-10)=-9$	11
A(P1,P7)	$22+0=22$	$12+(-10)=2$	$30+(-10)=20$	22
A(P4,P6)	$1+0.9=1.9$	$11+(-10)=1$	$12+(-10)=2$	2
A(P4,P5)	$12+0=12$	$2+(-10)=-8$	$22+(-10)=12$	12

i / j		j=0	j=1	j=2	j=3	j=4	j=5
			P5	P4	P1	P6	P7
i = 0		0	10	20	30	40	50
i = 1	P1	10	0	10	21	30	40
i = 2	P6	20	10	0	11	22	30
i = 3	P1	30	20	10	1	12	22
i = 4	P4	40	30	21	11	2	12

$A(m,n)$ i.e., $A(P4,P5) = 12$ is the optimal alignment for sessions S3 and S4.

Session similarity = Optimal alignment score/length of longer session = $12/5 = 2.4$