

24/3/24

Hizalama Algoritmaları

Global Hizalama → İki dizinin gerekirse boşluklar kullanılarak eşit uzunlukta olacak biçimde hizalanması

Yerel Hizalama → Dizilerin sadece belirli bir kısmının hizalanması

Matematiksel Tanım:

$x = x_1, x_2, \dots, x_m$, $y = y_1, y_2, \dots, y_n$ şekilde iki dizi olu.

hizalama boşlukların x indisinde $0, \dots, m$ ve y indisi için $0, \dots, n$ indislerine her iki dizideki harflerin karşısında bir harf veya bir boşluk bulunacak biçimde atamasıdır.

- Global hizalamada bu işlem tüm diziyi kapsayacak şekilde yapılırken yerel hizalamada belirli bir bölgeyi kapsayacak biçimde gerçekleştirilir.
- PAM, BLOSUM matrisleri → Proteinler için skorlama matrisleri.

PAM matrisi

- Global hizalama ile olmuştur.
- Az miktarda veri ile oluşturulmuştur.
- Global hizalamalarda ve evrimsel olarak yakın homologlarda daha iyi sonuç verir.
- Yüksek numaralı versiyonları daha az benzeyen dizileri hizalamak için optimize edilmiştir.

BLOSUM Matrisi

- Yerel hizalamalar ile oluşturulmuştur.
- Büyük miktarda veri ile oluşturulmuştur.
- Yerel hizalamalarda daha başarılıdır. Evrimsel olarak uzak homologlarda daha iyi sonuç verir.
- Düşük numaralı versiyonları daha az benzeyen dizileri hizalamak için optimize edilmiştir.

Needelman - Wunsch ile Global Hizalama

SEND, AND iki dizi:

	S	E	N	D	
A	$C(1,1)$	$C(1,2)$	$C(1,3)$	$C(1,4)$	$C(1,5)$
N	$C(2,1)$	$C(2,2)$	$C(2,3)$	$C(2,4)$	$C(2,5)$
D	$C(3,1)$	$C(3,2)$	$C(3,3)$	$C(3,4)$	$C(3,5)$
	$C(4,1)$	$C(4,2)$	$C(4,3)$	$C(4,4)$	$C(4,5)$

$s(a,b)$ → a ve b dizi elemanlarının hizalanmasından elde edilen skor

$$C(i,1) = \sum s(A(k), -), 1 \leq k \leq i$$

$$C(1,j) = \sum s(-, B(k)), 1 \leq k \leq j$$

→ Her açılan boşluk için -10 p. verişsek tablonun ilk satır ve sütunları şu şekilde olur:

	S	E	N	D
0	-10	-20	-30	-40
A	-10			
N	-20			
D	-30			

→ Skor kayıtlarının yanı sıra hangi hüreden - hürelere geldiğimizi de kaydetmemiz gerekir.

Her bir hücre için 3 alternatif var:

İlk Pozisyon	Skor	Kalın Dizi
S A	1	END ND
- A	-10	SEND ND
S -	-10	END AND

Gesleşme +1, boşluk -10

Bu üç alternatifi tabloda şu şekilde hesaplarız:

$$q_{\text{diag}} = C(i-1, j-1) + S(i, j)$$

$$q_{\text{up}} = C(i-1, j) + g$$

$$q_{\text{left}} = C(i, j-1) + g$$

Yerel Hizalama

Negatif skor için eşleşme : 0 p

→ Dizi içindeki tüm negatif skorlara 0 p.

Formülize edilmiş hali :

$F(a,b)$ eşleşme skoru a.ü.

$$F(i,0) = 0$$

$$F(0,j) = 0$$

$$F(i,j) = \max \left[0, F(i,j-1) + s(-, B(j)), \right. \\ \left. F(i-1,j) + s(A(i), -), \right. \\ \left. F(i-1,j-1) + s(A(i), B(j)) \right]$$

	-	G	C	T	G	G	A	A	G	G	C	A	T
-	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4
G	-4	+1	-3	-2	-2	+3	-1	-5	0	+5	+1	-3	-2
C	-4	-3	+2	-2	-6	-1	-5	-9	-4	+1	+10	+6	+2
A	-4	-2	-2	-6	-10	-5	+4	+9	+5	+1	+6	+5	+11
G	-4	+1	+3	-6	-1	+4	-1	+5	+1	+9	+5	+11	+7
A	-4	-3	-2	-10	-5	0	+9	+5	+1	+5	+1	+10	+6
G	-4	-2	-11										
C	-4												
A	-4												
C	-4												
G	-4												

ÖRNEK

doğru eşleşme : +4
yanlış eşleşme : +2
boşluk : -1

EQLLKALEFKL
KVLEFGY

Gap → -4
match → +5
Mis match → -4

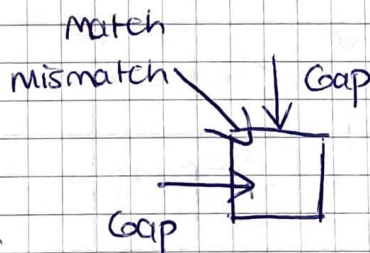
	-	E	Q	L	L	K	A	L	E	F	K	L
-	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	4	3	2	1	0	4	3
V	0	0	0	0	0	3	2	1	0	0	3	2
L	0	0	0	4	4	3	2	6	5	4	3	7
E	0	4	3	3	3	2	1	5	10	9	8	7
F	0	3	2	2	2	1	0	4	9	14	13	12
G	0	2	1	1	1	0	0	3	8	13	12	11
Y	0	1	0	0	0	0	0	2	7	12	11	10

	-	G	C	T	G	G	A	A	G	G	C	A	T
-	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	+5	+1	0	+5	+5	+1	0	+5	+5	+1	0	0
C	0	+1	+10	+6	+2	+1	+1	0	+1	+1	+10	+6	+2
A	0	0	+6	+6	+2	0	+6	+6	+2	0	+6	+15	+11
G	0	+5	+2	+2	+11	+7	+3	+2	+11	+7	+3	+11	+11
A	0	+1	+1	0	+7	+7	+12	+8	+7	+7	+3	+8	+7
G	0	+5	+1	0	+5	+12	+8	+8	+13	+12	+8	+4	+4
C	0	+1	+10	+6	+2	+8	+8	+4	+9	+9	+17	+13	+9
A	0	0	+6	+6	+2	+4	+13	+13	+9	+5	+13	+22	+18
C	0	0	+5	+2	+2	0	+9	+9	+9	+5	+10	+18	+18
G	0	+5	+1	+1	+7	+7	+5	+5	+14	+14	+10	+14	+14

1) Negatif deger duz

2) ~~Matris~~ Satir sutun 0

3) Localde en buyuk deger
alnr.



G A A G - G C A
G C A G A G C A

+22

+5
-4
-4

-4

Gap: -4

Match: +5

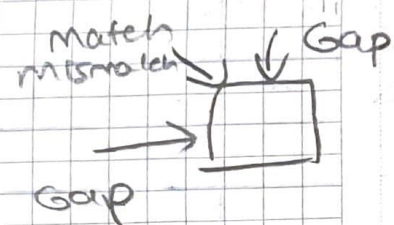
Mismatch: -4

	-	Y	E	C	H	Q	C	G	K	A	F	R
-	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	+5	+1	0	0	0	0	0	0	0	0	0
E	0	+1	+10	+6	+2	0	0	0	0	0	0	0
C	0	0	+6	+15	+11	+7	+5	+1	0	0	0	0
H	0	0	+2	+11	+20	+16	+12	+8	+4	0	0	0
Q	0	0	0	+7	+16	+25	+21	+17	+13	+9	+5	+1
C	0	0	0	+5	+12	+21	+30	+26	+22	+18	+14	+10
G	0	0	0	+1	+8	+17	+26	+35	+31	+27	+23	+19
Q	0	0	0	0	+4	+13	+22	+31	+31	+27	+23	+19
A	0	0	0	0	0	+9	+18	+27	+27	+36	+32	+28
F	0	0	0	0	0	+5	+14	+23	+23	+32	+41	+37
K	0	0	0	0	0	+1	+10	+19	+28	+28	+37	+37
T	0	0	0	0	0	0	+6	+15	+24	+24	+33	+33

Gap $\rightarrow -4$

match $\rightarrow +5$

Mismatch $\rightarrow -4$



1) Negatif deger olamaz

2) ilk satır

Y E C H Q C G K A F
Y E C H Q C G Q A F

	-	Y	E	C	H	Q	C	G	Q	A	F	K	T
-	0	-4	-8	-12	-16	-20	-24	-28	-32	-36	-40	-44	-48
Y	-4	15	+1	-3	-7	-11	-15	-19	-23	-27	-31	-35	-39
E	-8	+1	10	+6	+2	-2	-6	-10	-14	-18	-22	-26	-30
C	-12	-3	+2	15	+11	+7	+3	-1	-5	-9	-13	-17	-21
N	-16	-7	-7	+11	11	+7	+3	-1	-5	-9	-13	-17	-21
Q	-20	-11	-11	+7	+7	16	+12	+8	+4	0	-4	-8	-12
C	-24	-15	-15	-6	+3	+12	21	+17	+13	+9	+5	+1	-3
G	-28	-19	-19	-10	-1	+8	+17	26	+22	+18	+14	+10	+6
K	-32	-23	-23	-14	-5	+4	+13	+22	22	+18	+14	+19	+15
A	-36	-27	-27	-18	-9	0	+9	+18	+18	27	+23	+19	+15
F	-40	-31	-31	-22	-14	-4	+5	+14	+14	+23	32	+28	+24
R	-44	-36	-36	-26	-18	-8	+1	+10	+10	+19	+28	28	+24

match $\rightarrow +5$
 mismatch $\rightarrow -4$
 Gap $\rightarrow -4$

En büyük skor alınır.

Sağ alt köşeye başlanır.

CHQCGQAF
 CNQCGKAF

YEC HQCGQAFKT
 YEC NQCGKAFR