$\boxed{8/19}$ *look @ class notes for global alignment (needleman-wunsch) pseudocode & time complexity

BLOSUM MATRIX

alphabet: $\Sigma = \{A, B, C\}$

observed data: (could be, for instance, protein sequences from related species)

B	A	B	A
A	A	A	C
A	A	C	C
A	A	B	A
A	A	C	C
A	A	B	C

4 columns + 6 rows:

- 6 choose 2 ways to pair amino acids in a column
  - ↳ $\binom{6}{2} * 4$ columns $=$ $\boxed{\text{60 total pairings}}$
- 14 A's, 4 B's, 6 C's (24 total)
  - ⟶ frequency: $A \sim \frac{14}{24}$, $B \sim \frac{4}{24}$, $C \sim \frac{6}{24}$

to calculate this, go through each column of the observed data and calculate the # of ways to get the desired pair

aligned pairs	observed frequency	expected frequency	log likelihood ratio	rounded log likelihood ratio
A → A	$\left(\binom{5}{2}+\binom{6}{2}+\binom{2}{2}\right)/60 = \frac{24}{60}$	$\left(\frac{14}{24}\right)\left(\frac{14}{24}\right) = \frac{196}{576}$	$2\log\left(\frac{obs}{exp}\right) = 0.7$	1
A → B	$\frac{5+3}{60} = \frac{8}{60}$	$2\left(\frac{14}{24}\right)\left(\frac{4}{24}\right) = \frac{112}{576}$	-1.09	-1
A → C	$\frac{2+8}{60} = \frac{10}{60}$	$2\left(\frac{14}{24}\right)\left(\frac{6}{24}\right) = \frac{168}{576}$	-1.6	-2
B → B	$\binom{3}{2}/60 = \frac{3}{60}$	$\left(\frac{4}{24}\right)\left(\frac{4}{24}\right) = \frac{16}{576}$	-1.7	2
B → C	$\frac{6}{60}$	$2\left(\frac{4}{24}\right)\left(\frac{6}{24}\right) = \frac{48}{576}$	0.53	1
C → C	$\frac{\binom{3}{2}+\binom{4}{2}}{60} = \frac{7}{60}$	$\left(\frac{6}{24}\right)\left(\frac{6}{24}\right) = \frac{36}{576}$	1.8	2

can use these values to populate a biologically significant scoring scheme

$$\begin{array}{c|ccc} & A & B & C \\ \hline A & 1 & -1 & -2 \\ B & -1 & 2 & 1 \\ C & -2 & 1 & 2 \end{array}$$

Likelihood & Heuristic interpretation of "alignment score"

ex) X = AACGTAA
Y = ACGGTGT

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

4 matches, 2 indels, 2 mismatches

p = probabilty of mismatch
q = prob of mismatch
r = prob of indel

→ probability of alignment $= \mathbb{P}_A = p^4 q^2 r^2$

$$\Delta' = \log(\mathbb{P}_A) = 4\log(p) + 2\log(q) + 2\log(r)$$

$$S = \Delta' - 8\log(k)$$

k is a constant such that $\log\left(\frac{p}{k}\right) = 1$

$$S = \Delta' - 8\log k = 4\log p + 2\log q + 2\log r - 8\log k$$

$$= 4\log p - 4\log k + 2\log q - 2\log k + 2\log r - 2\log k$$

$$= 4\cancel{\log\left(\frac{p}{k}\right)}^{1} + 2\underbrace{\log\left(\frac{q}{k}\right)}_{-M} + 2\underbrace{\log\left(\frac{r}{k}\right)}_{-T}$$

$$= 4 - 2M - 2T$$

Thus, to maximize the likelihood of alignment, you have to maximize the alignment score