5 Exercises – Evaluation of alignment scores

1. **Non-parametric test**

   A non-parametric test is used to determine a p-value for the optimal score of a global alignment. Assume we randomly generated 9 sequences and calculated the alignment scores as follows.

   ```
   q: AACG
   ```

Seq No.	1	2	3	4	5	6	7	8	9
Score	0.2	0.4	0.5	1.2	1.2	1.5	1.9	2.2	2.1

   (a) What are H_0 (null hypothesis) and H_1 (alternative hypothesis) if you want to use a statistical hypothesis test to evaluate a global pairwise alignment in terms of finding homologues?

   (b) Calculate the p-value for the alignment below.
   ```
   q: AACG
   d: AGTG
   ```
   ```
   Score: 2
   ```
   The p-value can be calculated as:
   $$p = (b + 1)/(n + 1)$$
   where b is the number of randomly generated scores above the score of the original alignment, and n is the sample size.

   (c) Is the test result statistically significant when $\alpha = 0.05$?

   (d) What is the conclusion of the test in terms of finding homologues?

2. **Gumbel distribution**

Use the alignment scores between q and 5 randomly generated sequences below to answer the following questions.

```
q: ACGTA
```

Randomly generated sequence	Alignment score
ACGTA	4
TTACG	5
CGCGA	6
ATTAT	4
CGATC	6

(a) What is the mean of the scores?

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

(b) What is the standard deviation of the scores?

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

(c) The parameters μ and λ of the Gumbel distribution can be estimated from the mean and the variance of a sample. Calculate lambda and mu.

$$\text{lamda} \approx \frac{1.282}{s}$$

$$\text{mu} \approx \bar{x} - \frac{0.577}{\text{lamda}}$$

(d) What is the p-value when the score of an alignment is 4.55?

x	exp(x)
0	1
-1	0.3679
-2	0.1353
-3	0.0497

$$P[Y > 4.55] = 1 - F_Y(4.55) = 1 - \exp(-e^{-\lambda(4.55-\mu)})$$

(e) What is the conclusion of the test in terms of finding homologues?

3. **Bit score**

BLAST reports bit scores, which are equivalent to the information content that are calculated from raw scores.

- Bit score: $S' = \dfrac{(\lambda S - \ln K)}{\ln 2}$
- S: raw score
- K and λ: Karlin-Altschul statistics

x	$\ln x$
1	0
2	0.693
3	1.099

Assume K: 3 and λ: 0.1 and use the table above to answer the following questions.

(a) What is the bit score when the raw score is 80.29?

(b) $2^{S'}$ indicates the expected search space size that one can find an alignment with score at least S by chance alone. What is the $2^{S'}$ value when the raw score is 17.92?

4. **Bit score to e-value**

The e-value represents the expected number of hits when homologous sequence are searched on a database of a particular size. It can be calculated from a bit score as follows.

$$E = \frac{\text{Size of search space}}{2^{\text{bit}-\text{score}}}$$

(a) What is the e-value when the size of search space is 3200 and the bit score is 4?

(b) Assume we can calculate a search space size as $m \times n$ where m is the query sequence size and n is the total character size of a database. What is the e-value when m is 20, n is 4000, and the bit score is 3?

5. **Raw score to e-value** The e-value can be directly caudated from a raw score as $E(S) = Kmne^{-\lambda S}$. Use the values below to answer the following questions.

- K: 2
- λ: 0.1

x	exp(x)
0	1
-1	0.3679
-10	0.000045

(a) What is E(10) when m is 10 and n is 100?

(b) What is E(100) when m is 10 and n is 100?

(c) What is E(100) when m is 100 and n is 1000?