INF281 Exercise 05

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?

Solution: H_0: Sequences are not homologous, H_1: Sequences are homologous

(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.

Solution: p-value: 0.3

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

Solution: No.

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

Solution: q and d are not homologous.

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?

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

Solution: 5

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

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

Solution: 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.

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

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

Solution: lamda: 1.282, mu: 4.55

(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

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

Solution: $1 - 0.3679 = 0.6321$

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

Solution: The sequences of the alignment with score 4.55 are not homologous.

3. **Bit score**

BLAST reports bit scores, which are 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?

Solution: $\dfrac{0.1 \times 80.29 - 1.099}{0.693} = \dfrac{6.93}{0.693} = 10$

(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?

Solution: $S' = \dfrac{0.1 \times 17.92 - 1.099}{0.693} = \dfrac{0.693}{0.693} = 1; \quad 2^1 = 2$

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?

   Solution: $\frac{3200}{2^4} = \frac{3200}{16} = 200$

   (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?

   Solution: $\frac{20 \times 4000}{2^3} = \frac{80000}{8} = 10000$

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?

   Solution: $E(10) = 210100e^{-1} = 7358$

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

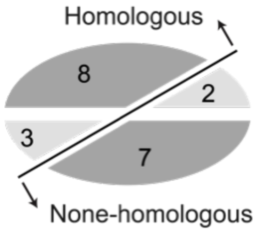
   Solution: $E(100) = 210100e^{-10} = 0.09$

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

   Solution: $E(100) = 21001000e^{-10} = 9$

6. **Basic measures from confusion matrix**

The oval representation below shows that a model has classified a test data set with 10 positives and 10 negatives and produced four outcomes.



(a) Fill each blank cell with one of the four classification outcomes - TP, FP, TN, and FN.

		Test data	
		Homologous	Non-homologous
Model classification	Homologous	TP	FP
	Non-homologous	FN	TN

(b) Make a confusion matrix for the oval representation.

		Test data	
		Homologous	Non-homologous
Model classification	Homologous	8	3
	Non-homologous	2	7

(c) Calculate the following basic evaluation measures for the oval representation. Round off the answer to two decimal places if necessary.

$$\text{Accuracy} = \frac{TP+TN}{P+N} = \frac{15}{20} = 0.75$$

$$\text{Error rate} = \frac{FP+FN}{P+N} = \frac{5}{20} = 0.25$$

$$\text{Sensitivity} = \frac{TP}{P} = \frac{8}{10} = 0.8$$

$$\text{Specificity} = \frac{TN}{N} = \frac{7}{10} = 0.7$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{8}{11} = 0.73$$

7. **Measures with multiple thresholds**

Create multiple confusion matrices by considering all possible threshold values. Assume that the test data set contains two positives and two negatives. The table below shows the scores given by a model that gives higher scores for the alignments with higher similarities.

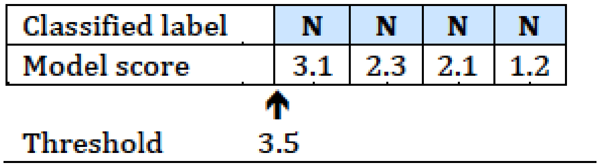
Test set label	P	P	N	N
Model score	2.1	3.1	2.3	1.2

(a) Fill the labels that match the sorted scores

Test set label				
Model score	2.1	3.1	2.3	1.2

(b) Fill the labels predicted by different threshold values.

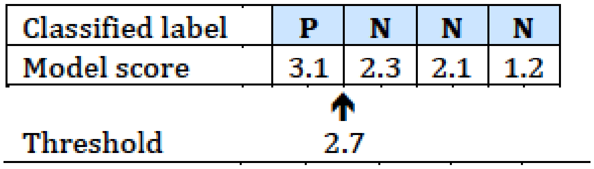
1. 3.5



Classified label	N	N	N	N
Model score	3.1	2.3	2.1	1.2

Threshold 3.5

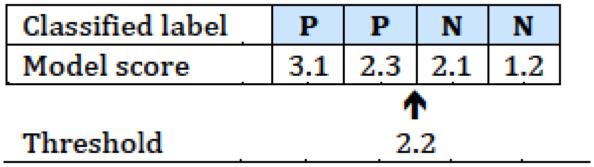
2. 2.7



Classified label	P	N	N	N
Model score	3.1	2.3	2.1	1.2

Threshold 2.7

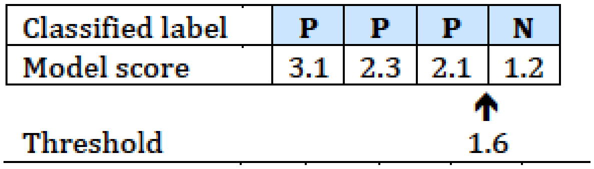
3. 2.2



Classified label	P	P	N	N
Model score	3.1	2.3	2.1	1.2

Threshold 2.2

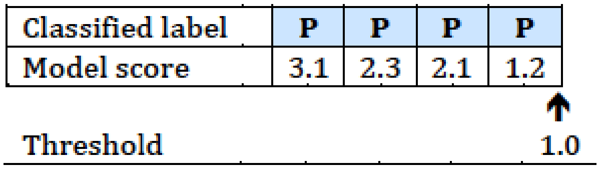
4. 1.6



Classified label	P	P	P	N
Model score	3.1	2.3	2.1	1.2

Threshold 1.6

5. 1.0



Classified label	P	P	P	P
Model score	3.1	2.3	2.1	1.2

Threshold 1.0

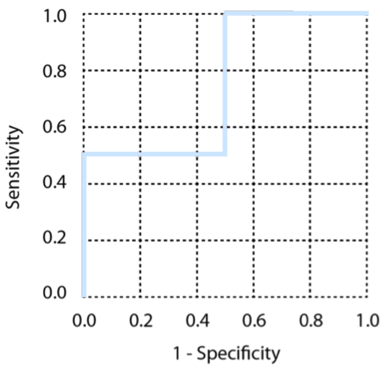
(c) Use the labels in (a) and (b) and calculate TP, FP, TN, and FN for all threshold values.

Threshold	TP	FP	TN	FN
3.5	0	0	2	2
2.7	1	0	2	1
2.2	1	1	1	1
1.6	2	1	1	0
1	2	2	0	0

(d) Use the result in (c) and calculate basic evaluation measures. Round off the answer to one decimal place if necessary.

Threshold	Specificity	1 - specificity	Sensitivity
3.5	1	0	0
2.7	0.5	0.5	0.5
2.2	0.5	0.5	0.5
1.6	0	1	1
1	0	1	1

(e) Draw a ROC curve for the calculated evaluation measures in (d).



(f) Calculate the area under the curve of the curve in (e).

Solution: 0.75

(g) Evaluate the ROC curve in your own words.

Solution: The model performs better than random classifiers