Wintersemester 2021/22

Xiheng He

Lisanne Friedrich

Exercises for Algorithmic Bioinformatics II

Assignment 5

Xiheng He

November 2021

Exercise 3 (BLAST Expectation Values, 10P):

Consider $\lambda = 0.32, K = 0.137, H = 0.401$ and a database of 22 million amino-acids. The best hit of length 103 in a query sequence of length 112 has score $s = 65$.

Use the expected length of an average HSP for edge-correction and do not use the sum statistics.

(a) Calculate the E-value

$$E = Kmne^{-\lambda \cdot S} = 0.137 \cdot 103 \cdot 112 \cdot e^{-0.32 \cdot 65} = 1.463695 \times 10^{-6}$$

(b) Calculate the E-value after *Edge-Correction*.

$$\begin{aligned}
\because H \approx \frac{\lambda\mu}{E(l)} = \frac{\ln(Kmn)}{E(l)} &\Longrightarrow E(l) \approx \frac{\lambda\mu}{H} = \frac{\ln Kmn}{H} \\
&= \frac{\ln(0.137 \cdot 103 \cdot 112)}{0.401} = 18.3677 \\
\therefore N_1' &= N_1 - E(l) = 103 - 18.3677 = 84.63 \approx 85 \\
N_2' &= N_2 - E(l) = 112 - 18.3677 = 93.6323 \approx 94 \\
E' &= N_1' \cdot N_2' \cdot K \cdot e^{-\lambda \cdot S} = 85 \cdot 94 \cdot 0.137 \cdot e^{-0.32 \cdot 65} = 1.013776 \times 10^{-6}
\end{aligned}$$

(c) Calculate the E-value after *Multiple-Testing-Correction.*

$$E'(r=1) = 2 \cdot E = 2.92739 \times 10^{-6}$$

$$P \approx 1 - e^{-E'} = 1 - e^{-2.92739 \times 10^{-6}} = 2.92739 \times 10^{-6}$$

$$\Longrightarrow E''_{MTC} = P \cdot \frac{D}{N_2} = 2.92739 \times 10^{-6} \cdot \frac{22 \times 10^6}{103} = 0.625267767 \approx 0.625$$

(d) Calculate the E-value after *Edge-Correction* and *Multiple-Testing-Correction.*

$$E'(r=1) = 2N_1'N_2'Ke^{-\lambda S} = 2 \cdot E_{ETC} = 2.027552 \times 10^{-6}$$

$$P \approx 1 - e^{-E'} = 1 - e^{-2.027552 \times 10^{-6}} = 2.0275 \times 10^{-6}$$

$$\Longrightarrow E''_{MTC} = P \cdot \frac{D}{N_2} = 2.0275 \times 10^{-6} \cdot \frac{22 \times 10^6}{103} = 0.4330689 \approx 0.433$$