Xiheng He

Lisanne Friedrich

Exercises for Algorithmic Bioinformatics II

Assignment 5

Xiheng He

November 2021

Exercise 4 (BLAST Statistics, 10P):

Use the given target distribution $q(i, j)$ and the frequencies $p(i)$ of a 2-nucleotide genome with $\lambda = 0.32$ to calculate

(a) all s_{ij} scores

(b) the *NATS* score

(c) the expected score of a nucleotide pair in a HSP

(d) the expected score of a random nucleotide pair in general

As always, please explain/justify your results.

q(i,j)	C	G
C	0.48	0.02
G	0.04	0.46

p(i)	C	G
	0.55	0.45

(a)

$$s_{ij} = (\ln(\frac{q_{ij}}{p_i \cdot p_j}) \cdot \frac{1}{\lambda})$$

$$\Longrightarrow$$

$$s_{cc} = \ln(\frac{0.48}{0.55 \cdot 0.55} \cdot \frac{1}{0.32}) = 1.4428$$
$$s_{cg} = \ln(\frac{0.02}{0.55 \cdot 0.45} \cdot \frac{1}{0.32}) = -7.8615$$
$$s_{gc} = \ln(\frac{0.04}{0.45 \cdot 0.55} \cdot \frac{1}{0.32}) = -5.6954$$
$$s_{gg} = \ln(\frac{0.46}{0.45 \cdot 0.45} \cdot \frac{1}{0.32}) = 2.5640$$

(b)

$$H(NATS) = \lambda \sum_{i,j} s_{ij} \cdot q(i,j) = 0.32 \times (0.48 \times 1.4428 + 0.02 \times -7.8615 + 0.04 \times -5.6954 + 0.46 \times 2.5640)) = 0.4758$$

(c)

$$E_{HSP} = \frac{H}{\lambda} = \frac{0.4758}{0.32} = 1.4869$$

(d)

$$E = \sum_{i,j} p_i p_j s_{ij} = 0.55 \times 0.55 \times 1.4428 + 0.45 \times 0.55 \times -7.8615 + 0.45 \times 0.55 \times -5.6954 + 0.45 \times 0.45 \times 2.5640 = -2.3997$$