Präsenzübungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, WS 2012/2013
Prof. Dr. Jens Stoye · Nina Luhmann · Linda Sundermann

http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2012winter/SequenzAnalyse

Präsenzübungsblatt 1, Woche 43/2012

Aufgabe 1 (Sequenzwahrscheinlichkeiten)

1. Wie wahrscheinlich ist es, die DNA-Sequenz $s = TATAAA$ per Zufall zu erhalten? Nimm dabei an, dass jede Base mit der selben Wahrscheinlichkeit auftritt. Gib auch die Formel an, die du zur Berechnung benutzt hast.
2. Wieso wird diese Sequenz nicht nur aufgrund von Zufall in einem Eukaryoten-Genom auftauchen?
3. Gegeben sei nun eine weitere DNA-Sequenz $s2$ der Länge 7. Wie wahrscheinlich ist es, dass genau 3 'As' darin vor kommen? Gebe auch hierzu eine Formel an.
4. Wie wahrscheinlich ist es, dass in der gleichen Sequenz $s2$ mindestens ein 'G' vorkommt?

Aufgabe 2 (Anzahl von Subsequenzen)

1. Wie viele Subsequenzen der festen Länge k hat ein String der Länge n?
2. Wie viele Subsequenzen mit den Längen $1 \leq k \leq n$ hat ein Wort der Länge n insgesamt?

Versuche, dir die Formeln jeweils anschaulich klar zu machen.

Aufgabe 3 (Verschiedene Metriken)

Aus der Vorlesung kennst du bereits folgende Metriken mit ihren Definitionen:

$$d_1(x,y) := \sum_{i=1}^{n} |x_i - y_i| \qquad \text{(Mannheim metric oder Manhattan distance)}$$

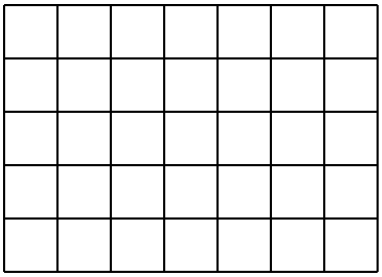
$$d_2(x,y) := \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \qquad \text{(Euclidean distance)}$$

$$d_\infty(x,y) := \max_{i=1,\ldots,n} |x_i - y_i| \qquad \text{(maximum metric)}$$

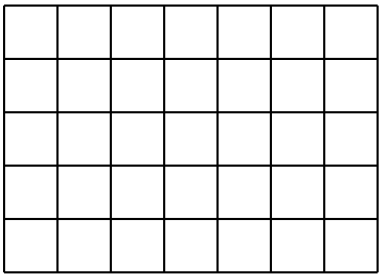
$$d_H(x,y) := \sum_{i=1}^{n} \mathbb{1}_{x_i \neq y_i} \qquad \text{(Hamming distance)}$$

Zeiche in die folgenden Koordinatensysteme die Punkte A und B jeweils so ein, dass sich die folgenden Distanzen ergeben:

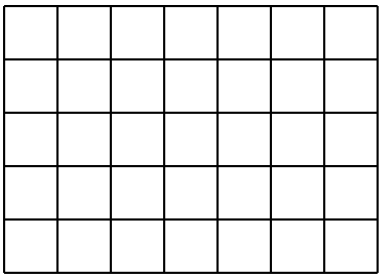
1. $d_1(A,B) = 9$



2. $d_2(A, B) = \sqrt{34}$



3. $d_\infty(A, B) = 3$



4. $d_H(A, B) = 2$