

## Algorithmische Bioinformatik Übungsblatt 11

Ausgabe: 28. Januar 2020 · Besprechung: keine

**Aufgabe 11.1** (Metriken) Gegeben ist folgende symmetrische Distanzmatrix  $D$  auf 5 Objekten. Handelt es sich jeweils um eine Metrik, additive Metrik, Ultrametrik? Gib jeweils ein Zertifikat bzw. einen Baum an.

$$D = \begin{pmatrix} 0 & & & & \\ 9 & 0 & & & \\ 7 & 5 & 0 & & \\ 9 & 5 & 5 & 0 & \\ 7 & 9 & 7 & 9 & 0 \end{pmatrix}$$

**Aufgabe 11.2** Gegeben ist  $D(a, b) = 20$ ,  $D(a, c) = 15$ ,  $D(b, c) = 30$ . Berechne die Kantenlängen des ungewurzelten Baums mit den Blättern  $a, b, c$ .

**Aufgabe 11.3** (Agglomeratives Clustern auf additiven Metriken) Wenn die Eingabedistanzen zu UPGMA (oder einem anderen agglomerativen Verfahren) eine additive Metrik bilden, kann eine falsche Baumtopologie erzeugt werden: Betrachte das Beispiel eines Quartetts mit Blattkantenlängen  $(A:1, B:3) - 1 - (C:1, D:3)$ .

Wie sieht die Distanzmatrix aus, und was tun jeweils Single Linkage Clustering, Complete Linkage Clustering, WPGMA, UPGMA?

**Aufgabe 11.4** Beweise die Aussage aus dem Skript, dass die Maximierung von  $Z_D(u, v)$  äquivalent ist zur Minimierung von  $S_D(u, v)$  (definiert in (7.1)).

**Aufgabe 11.5** Beweise das Verschiebungslemma.

**Aufgabe 11.6** (\*) UPGMA (und verwandte Verfahren) fassen immer *das Objektpaar mit kleinstem Abstand* zu einem neuen Objekt zusammen. Lässt sich auch für NJ das Auswahlkriterium in einem einfachen Satz anschaulich formulieren?

**Aufgabe 11.7** (schnelles UPGMA, \*) Lies das im Skript angegebene Paper von Gronau und Moran (insbes. Abschnitt 2; Zugriff ist über die Rechner von der TU Dortmund möglich) und implementiere das exakte UPGMA-Verfahren in quadratischer Zeit mit Hilfe der dort erwähnten NN-Ketten (NN-chains).