

4. Übung zur Vorlesung “Bioinformatische Methoden in der Genomforschung”

Sebastian Böcker, Martin Hoffmann

Ausgabe: 16.11.2021

Abgabe: 22.11.2021

Aufgabe 1 (5 Punkte)

Erstellen Sie mit Hilfe von in Aufgabe 3 der letzten Übungsserie berechneten Distanzmatrizen jeweils zwei hierarchische Cluster-Bäume: einer mit *UPGMA* (*Unweighted Pair Group Method with Arithmetic mean*) und anderer mit *Single Linkage* als Intercluster-Distanz. Wie unterscheiden sich die Ergebnisse? Welche sind “besser”?

Aufgabe 2 (5 Punkte)

Zeigen Sie, dass *UPGMA* tatsächlich “unweighted” ist.

Aufgabe 3 (10 Punkte)

Gegeben ist eine Distanzmatrix D auf Objekten $X = \{1, \dots, n\}$ sowie ein hierarchisches Clustering in Form eines Wurzelbaums $T = (V, E)$. Um das hierarchische Clustering zu Visualisieren, sollen die Blätter des Baums so umsortiert werden, dass die Summe der Distanzen zwischen benachbarten Blättern im Baum minimiert wird.

Eine *lineare Anordnung* ist eine Permutation π der Zahlen $1, \dots, n$. Die lineare Anordnung π heißt *konsistent* mit T , wenn die Anordnung durch Umdrehen (Flip) von inneren Knoten von T erzeugt werden kann. Gesucht ist eine lineare Anordnung π , die konsistent mit T ist und die Zielfunktion

$$\sum_{j=1}^{n-1} D(\pi(j), \pi(j+1))$$

minimiert. Finden Sie einen Algorithmus mit Laufzeit $O(n^5)$, der das obige Problem löst. (Zusatzaufgabe: Zeigen Sie wenn möglich, dass ihr Algorithmus Laufzeit $O(n^4)$ hat, oder finden Sie einen schnelleren Algorithmus.)