3. Übung zur Vorlesung
"Bioinformatische Methoden in der Genomforschung"

Sebastian Böcker, Martin Hoffmann

Ausgabe: 09.11.2017
Abgabe: 23.11.2017

Aufgabe 1 (2 Punkte)

Was sind die Aufgaben der Microarray Datenanalyse? Welche Arbeitschritte werden häufig dafür gemacht?

Aufgabe 2 (2 Punkte)

Häufig werden bei der statistischen Analyse von Expressionsdaten Signifikanzen (p-values) berechnet. Vervollständigen Sie den folgenden Satz: "Ein p-value (Signifikanz) von 0.01 beim Test auf differentielle Genexpression bedeutet…".

Aufgabe 3 (6 Punkte)

Gegeben sei die Matrix der Genexpressionswerte gemessen für vier Gene an vier (aufeinanderfolgenden) Tagen:

	Tag 1	Tag 2	Tag 3	Tag 4
Gen 1	0.564	−0.038	−0.561	−1.315
Gen 2	0.606	0.621	−0.83	−1.681
Gen 3	−0.555	−0.224	0.673	0.78
Gen 4	0.238	−0.764	−1.371	−1.868

1. Berechnen Sie die Distanzmatrizen für folgende Distanzen:

   (a) Euklidische

   $$d_E(x_i, x_j) = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$$

   (b) Manhattan

   $$d_M(x_i, x_j) = \sum_k |x_{ik} - x_{jk}|$$

   (c) Korrelation

   $$d_C(x_i, x_j) = 1 - \frac{\sum_k ((x_{ik} - \mu_i)(x_{jk} - \mu_j))}{n\sigma_i\sigma_j}$$

   wobei x_{ik} der Expressionswert des i-ten Gens am k-ten Tag ist, μ_i der Mittelwert, σ_i die Standardabweichung der Expressionswerte des i-ten Gens und n die Anzahl der Tage ist.

   $$\sigma_i = \sqrt{\frac{1}{n}\sum_k (x_{ik} - \mu_i)^2}$$

2. Normalisieren Sie die Eingabematrix mit dem folgenden *Standardisierungsansatz*:

   Schritt 1. Für jeden Wert in der Zeile verwenden Sie die Transformation $x \mapsto \frac{x-\mu}{\sigma}$, wobei μ der Mittelwert und σ die Standardabweichung der Zeilenwerte ist.

   Schritt 2. Für jeden Wert in der Spalte verwenden Sie dieselbe Transformation mit dem Unterschied, dass der Mittelwert und die Standardabweichung nun über die Spalten berechnet werden.

   Wiederholen Sie den Schritt 1.

   Wiederholen Sie den Schritt 2.

   usw ...

   Wie sieht die Matrix nach einem, zwei, drei Schritte aus? (Zusatzaufgabe: Wie sieht die Matrix nach 10, 100, 1000 Schritten aus?) Was können Sie dabei beobachten?

Aufgabe 4 (5 Punkte)

Erstellen Sie mit Hilfe von in Aufgabe 3 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 5 (5 Punkte)

Zeigen Sie, dass *UPGMA* tatsächlich "unweighted" ist.

Aufgabe 6 (10 Punkte)

Gegeben ist eine Distanzmatrix D auf Objekten $X = \{1, ..., n\}$ 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 $1, ..., n$. 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.)