# 6. Übung zur Vorlesung "Algorithmische Massenspektrometrie"

Wintersemester 2020/2021

Sebastian Böcker, Kai Dührkop

Ausgabe: 09. Dezember 2020, Abgabe: 15. Dezember 2020

1. **Peak-Counting-Score:** Gegeben seien zwei Peaklisten $M = \{150, 180, 230, 310, 475\}$ und $M' = \{150, 190, 250, 315, 485\}$. Berechnen Sie den Peak-Counting-Score für $\delta = 5$, und $\delta = 10$.

   (1 Punkt)

2. **Alignment von Spektren:** Gegeben seien das gemessene Spektrum {200,300,500,515,700} und die beiden Referenzspektren {200,510,705,850} und {190,310,490,710}, sowie die Scoring-Funktion

$$\delta(m, m') = 2 - \frac{1}{5}|m - m'| \tag{1}$$

$$\delta(m, \varepsilon) = \delta(\varepsilon, m') = -1 \tag{2}$$

   Als Alignment zweier Spektren $M = m_1, m_2, \cdots, m_k$ und $M' = m'_1, m'_2, \cdots, m'_l$ bezeichnen wir eine Menge von (maximal k+l) Zuordnungen $(a, b) \in (M \cup \{\epsilon\}) \times (M' \cup \{\epsilon\})$ so dass:

   (a) Jeder Peak $m \in M$ sowie jeder Peak $m' \in M'$ exakt einmal in einer der Zuordnungen enthalten ist.

   (b) Es keine Zuordnung $(\epsilon, \epsilon)$ gibt.

   (c) Für alle Zuordnungen $(a, b)$ und $(x, y)$ gilt: $x > a \equiv y > b$. Optimale Alignments erfüllen diese Bedingung für sinnvolle Scoring Funktionen immer, daher spielt diese Bedingung für diese Aufgabe keine Rolle.

   Ein Beispiel für ein Alignment der beiden Referenzspektren zueinander wäre: $(200, 190)$, $(\epsilon, 310)$, $(510, 490)$, $(705, 710)$, $(850, \epsilon)$. Der Score eines Alignments ist die Summe der Scoring-Funktion über alle Zuordnungen. In diesem Beispiel: $\delta(200, 190) + \delta(\epsilon, 310) + \delta(510, 490) + \delta(850, \epsilon) = 0 - 1 - 2 - 1 = -4$.

   Als optimales Alignment bezeichnet man das Alignment mit maximalem Score.

   (a) Geben Sie die Rekurrenz für eine dynamische Programmierung an, die das optimale Alignment zweier Spektren für die gegebene Scoring-Funktion berechnet. Wie ist die Definition ihrer DP-Tabelle. Sie können sich am Needleman–Wunsch Algorithmus für Sequenzalignments zweier Strings orientieren.

   (b) Stellen Sie die zwei DP-Tabellen für die Alignments des gemessenen Spektrums gegen jeweils eins der Referenz-Spektren auf. Welches der beiden Referenzspektren ist dem gemessenem Spektrum am ähnlichsten?

   (6 Punkte)

3. **Statistisches Modell:** Das Scoring in Aufgabe 2 war sehr willkürlich festgelegt. Sinnvoller ist es, ein statistisches Modell für das Scoring zu verwenden und Log-Likelihoods oder Log-odds als Scores zu benutzen.

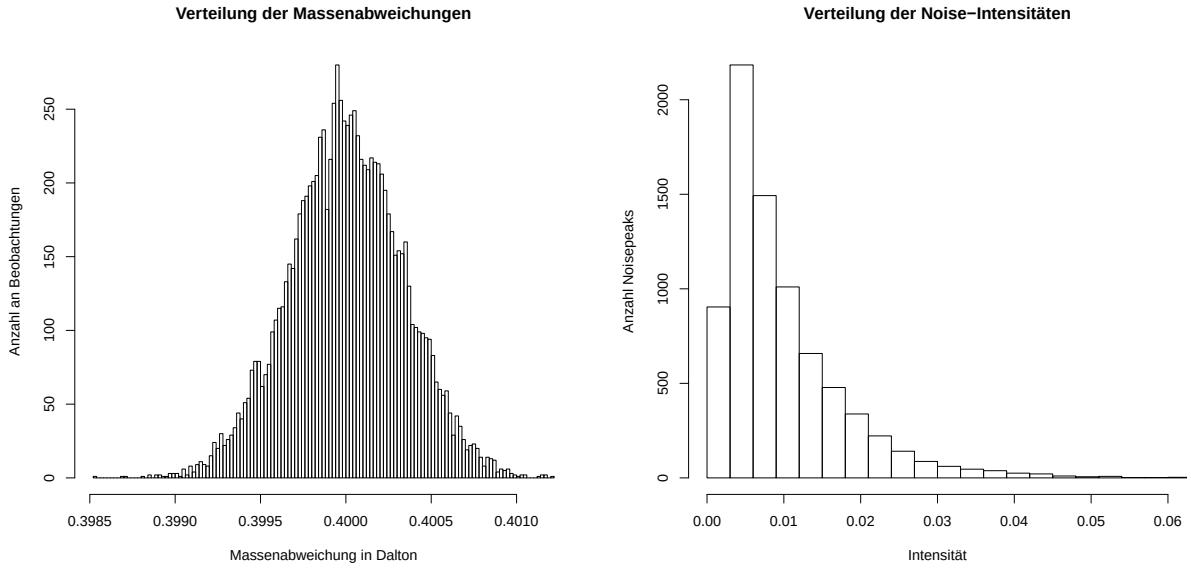


Abbildung 1: Das linke Histogramm zeigt die Verteilung der Massenabweichungen zwischen gemessenen Peaks und ihrer theoretischen Masse. Das rechte Histogramm zählt die Anzahl an Noisepeaks mit bestimmter Intensität.

(a) Warum verwendet man überhaupt logarithmierte (Wahrscheinlichkeits-)Werte? Was ist der Vorteil dabei?

(b) Wann immer wir einen Messfehler modellieren wollen, der durch eine Vielzahl von voneinander unabhängigen und zufälligen Prozessen entsteht, ist eine Normalverteilung eine gute Annahme. Warum ist dem so?

(2 Punkte)

4. **Wahrscheinlichkeitsverteilungen** Um das statistische Modell zu prüfen, betrachten wir viele Spektren von denen wir die Erklärung der Peaks kennen. Fig.1 zeigt ein Histogramm mit den Massenabweichungen zwischen den gemessenen Peaks und der theoretischen Masse der Compounds sowie ein Histogramm mit den Intensitäten aller Noise-Peaks.

(a) Im Histogramm ist zu sehen, dass die Massenabweichungen normalverteilt sind. Allerdings ist der Erwartungswert der Abweichung nicht 0. Welche Art von Fehler hat dies verursacht und was kann man tun um den Fehler aus seinen Daten herauszurechnen?

(b) Im zweiten Histogramm zeigen die Noise-Peaks ab einem bestimmten Intensitätstreshold eine Exponentialverteilung. Vor diesem Treshold hingegen nimmt die Zahl der Noisepeaks ab, statt exponentiell zuzunehmen. Wie ist das zu erklären? Macht es dennoch Sinn eine Exponentialverteilung der Noise-Peaks anzunehmen?

(2 Punkte)