6. Übung zur Vorlesung "Algorithmische Massenspektrometrie"

Wintersemester 2018/2019

Sebastian Böcker, Martin Hofmann

Ausgabe: 29. November 2018, Abgabe: 06. Dezember 2018

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}. Sei $\text{score}(m, m') = 2 - \frac{1}{5}|m - m'|$ die Scoring-Funktion aus der Vorlesung, und $\text{score}(m, \varepsilon) = \text{score}(\varepsilon, m') = -1$. Berechnen Sie durch dynamische Programmierung die Scores für die beiden Alignments des gemessenen Spektrums. Bestimmen Sie für den höheren Score das zugehörige Alignment. Geben Sie in der Lösung die Alignment-Tabelle und die optimalen Alignments mit an.

   (3 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.

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

   (b) Es ist immer sinnvoller einem intensiven Peak mehr zu trauen als einem weniger intensiven. Entsprechend sollte ein Spektralalignment, das viele intensive Peaks erklärt (matcht) besser bewertet werden als ein Alignment welches nur wenig intensive Peaks erklärt. Eine Möglichkeit das umzusetzen, wäre das Addieren der Log-Likelihoods, die aus der Massenabweichung berechnen wurden, mit der Intensität der erklärten Peaks. Warum macht dies statistisch sogar Sinn, bzw. wie lässt sich dies über ein statistisches Modell erklären?

   (c) 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?

   (4 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.

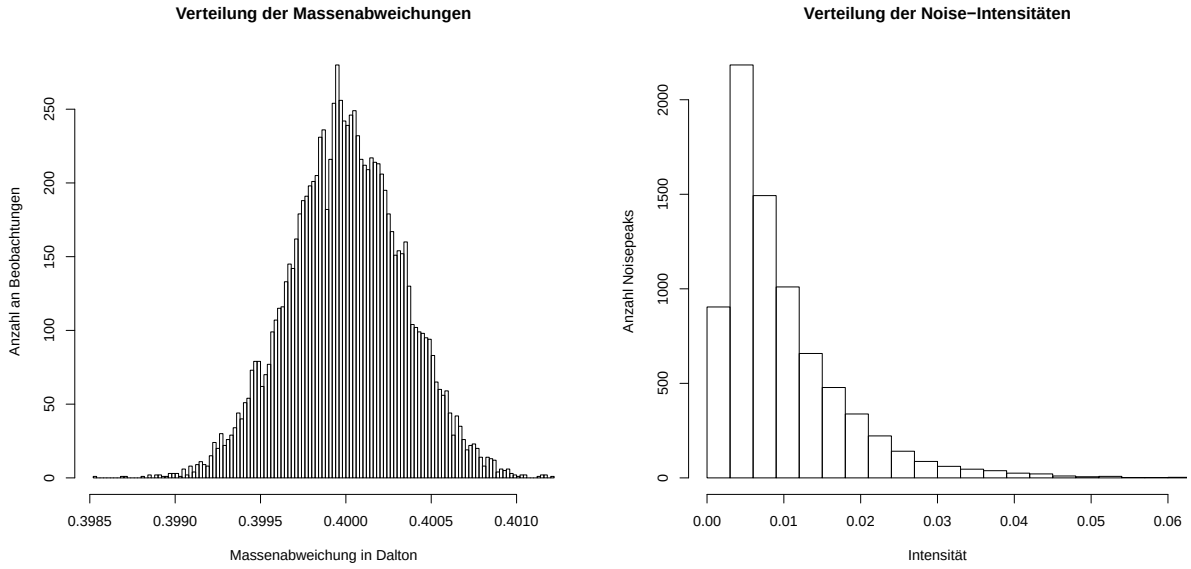

Figure 1: Das linke Histogramm zeigt die Verteilung der Massenabweichungen zwischen gemessenen Peaks und ihrer theoretischen Masse. Das rechte Histogramm zählt die Anzahl an Noisepeaks mit bestimmter Intensität. Beide Histogramme sind nicht aus realen Daten bestimmt, sondern lediglich simuliert.

(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?

(2 Punkte)