6. Übung zur Vorlesung "Algorithmische Massenspektrometrie"

Wintersemester 2014/2015

Sebastian Böcker, Kai Dührkop

Ausgabe: 05. Dezember 2014, Abgabe: 11. Dezember 2014

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 $\delta = 10$.

   (2 Punkt)

2. **Alignment von Spektren:** In der Praxis macht es Sinn kleine Messfehler besser zu bewerten als große Messfehler, statt eine maximal erlaubte Abweichung anzunehmen. Angenommen wir haben eine Scoring-Funktion $\delta : \mathbb{R}^2 \to \mathbb{R}$ die zwei Peak-Massen auf einen Score abbildet, mit: $\delta(m, m') := 5 - |m - m'|$. Schreiben Sie einen Algorithmus der zwei Spektren so miteinander aligniert (sprich: Paare von Peaks bildet), dass der Score dieser Paare maximiert wird. Hinweis: Das Problem lässt sich mittels dynamischer Programmierung in quadratischer Zeit lösen.
   Berechnen sie das optimale Alignment für die Spektren $M_1 = \{200, 203, 301, 350\}$ und $M_2 = \{204, 300, 303, 400\}$.

   (4 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? Hinweis: Noise-Intensitäten verhalten sich ungefähr exponentialverteilt.

   (c) Neben der Exponentialverteilung ist auch die Paretoverteilung eine Möglichkeit, Noise-Intensitäten zu modellieren. Worin unterscheidet sich die Paretoverteilung von der Exponentialverteilung (Hinweis: *Heavy-tailed distribution*)? Warum macht die Paretoverteilung möglicherweise mehr Sinn.

   (d) 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?

   (6 Punkte)

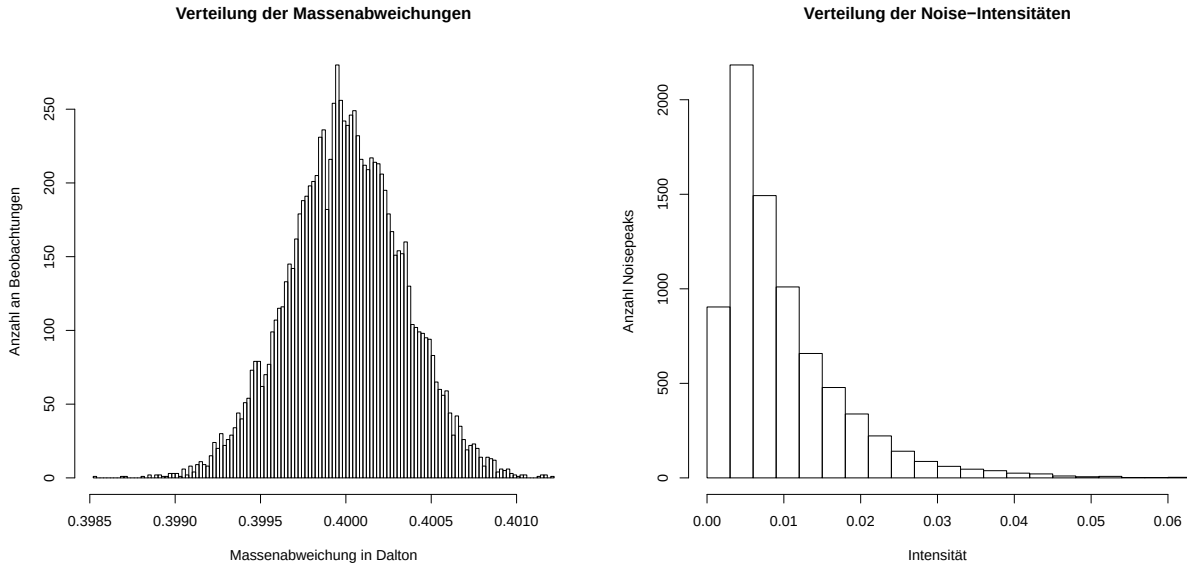

Abbildung 1: Das linke Histogramm zeigt die Verteilung der Massenabweichungen zwischen gemessenen Peaks und ihrer theoretischen Masse. Das rechte Histogramm zählt die Anzahl an Noisepeaks mit bestimmter Intensität. Beide Histogramme sind nicht aus realen Daten bestimmt, sondern lediglich simuliert.

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 Intensitätsthreshold eine Exponentialverteilung. Vor diesem Threshold hingegen nimmt die Zahl der Noisepeaks ab, statt exponentiell zuzunehmen. Wie ist das zu erklären?

(4 Punkte)