7. Übung zur Vorlesung "Algorithmische Massenspektrometrie"

Wintersemester 2020/2021

Sebastian Böcker, Kai Dührkop

Ausgabe: 16. Dezember 2020, Abgabe: 07. Januar 2021

1. **Strings-Zufallsgenerator:**

   Vom vierten Übungszettel wissen wir, dass die Anzahl der Strings mit Masse M und beliebiger Länge durch die Rekurrenz

   $$D[M] = \sum_{\sigma \in \Sigma, \mu(\sigma) \leq M} D[M - \mu(\sigma)]$$

   mit $D[0] = 1$ berechnet werden kann. Beschreiben Sie einen Algorithmus, der zufällig einen String mit gegebener Masse $M \in \mathbb{N}$ in Zeit $O(M/a_1)$ zieht.

   (5 Punkte)

2. **Strings-Zufallsgenerator mit unterschiedlich wahrscheinlichen Buchstaben:**

   Modifizieren Sie den Algorithmus aus Aufgabe 1 für den Fall, dass die Buchstaben $\sigma \in \Sigma$ nicht gleichwahrscheinlich sind, sondern wir für jeden Buchstaben σ eine Wahrscheinlichkeit $p(\sigma)$, $\sum_{\sigma \in \Sigma} p(\sigma) = 1$ gegeben haben. Achtung: Hierfür muss die Berechnung der DP Tabelle D leicht verändert werden!

   (4 Punkte)

3. **Empirische Verteilungen schätzen**

   Angenommen wir hätten einen Algorithmus, der ein Peptid-Massenspektrum in einer Datenbank von simulierten Peptidspektren sucht. Mittels des Algorithmus aus Aufgabe 1 können wir für eine Datenbankanfrage (gemessenes Peptidspektrum mit Muttermasse M) eine beliebige Anzahl zufälliger Peptide mit Muttermasse M generieren. In Übungsserie 1 hatten wir einen Algorithmus gebaut, der für eine gegebene Peptidsequenz ein theoretisches Massenspektrum simuliert. Beides zusammen erlaubt es uns, eine Datenbank von Zufallspeptiden zu bauen und Scores zwischen den gemessenen Peptiden und den Zufallspeptiden zu berechnen. Abbildung 1 zeigt die Verteilung dieser Scores.

   (a) Um was für eine Verteilung könnte es sich handeln? (Spoiler: Schauen Sie sich die folgenden Verteilungen an Normalverteilung, Lognormalverteilung, Exponentialverteilung, Paretoverteilung, Extremwertverteilung, Laplaceverteilung.)

   (1 Punkt)

   Wir suchen eine Peptidsequenz in unserer biologischen Datenbank: unter den 200 Kandidaten mit gleicher Muttermasse liefert der beste Treffer einen Score von 25. Wir wollen nun die Signifikanz dieses Treffers berechnen.

(b) In *table_1.txt* finden Sie die Scores beim Suchen unserer Peptidsequenz in einer Datenbank von Zufallspeptiden gleicher Masse. schätzen Sie aus diesen Scores die Parameter ihrer Verteilung.

(4 Punkte)

(c) Berechnen sie den p-Value und den korrigierten p-Value für den Score 25.

(3 Punkte)

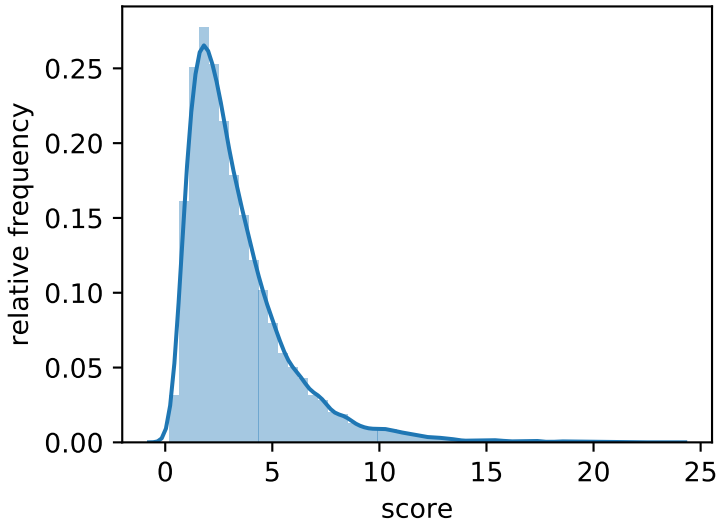

Abbildung 1: Verteilung der Scores zwischen gemessenen Peptiden und zufälligen Peptiden gleicher Masse.