

4. Übung Skriptsprachen in der Bioinformatik

Sommersemester 2015

Kai Dührkop

Ausgabe: 24.09.2015

Die Datei *compound.mgf* enthält ein Tandem-Massenspektrum im MGF (Mascot-Generic) Format. Mit einem Massenspektrometer werden die Massen von Molekülen einer Probe detektiert. Ein solches Spektrum weist jeder Masse eine Intensität zu und damit eine ungefähre relative Anzahl an Molekülen, die mit dieser Masse gemessen wurden. Relevant für diese Aufgabe sind lediglich die Werte-Paare von Masse und Intensität, andere Metainformationen können ignoriert werden. Diese Werte-Paare werden als **Peaks** bezeichnet. Bei einem Tandem-Massenspektrum wird ein Molekül selektiert und fragmentiert und zusammen mit diesen Fragmenten gemessen.

Aufgabe 1 Schreiben Sie eine Funktion, die ein MGF File wie *compound.mgf* einliest und einen zweidimensionalen Numpy-Array zurückgibt. Jede Zeile des Arrays sollte zwei Werte (Masse und Intensität) enthalten.

Aufgabe 2 Schreiben sie eine Funktion **baseline**, die ein Numpy-Array wie oben beschrieben als Eingabe bekommt, sowie einen Threshold als zweiten Parameter. Die Funktion soll einen neuen Array zurückgeben, in dem nur noch Peaks (also Zeilen) enthalten sind, deren Intensität über dem gegebenen Threshold liegt.

Aufgabe 3 Das gemessene, intakte Ion hat eine Masse von ungefähr 448 Da. Zusätzlich wurden mehrere Fragmente (Bruchstücke) dieses Ions gemessen. Alle anderen Peaks in dem Spektrum sind Noise. Da das intakte Ion eine Masse von 448 Da hat, müssen alle Massen über 450 Da zwangsläufig Noise sein. Berechnen sie den Median, Mittelwert und die 1 % 25 % und 99 % Percentile der Peak-Intensitäten jeweils für alle Peaks unter 450 Da und für alle Peaks über 450 Da und schreiben sie die unterschiedlichen Werte in die Ausgabe.

Aufgabe 4 Noise-Intensitäten lassen sich über eine Exponential- oder Paretoverteilung beschreiben. Fitten Sie jeweils die Parameter der Exponentialverteilung und der Paretoverteilung an den Intensitäten der Peaks mit Masse über 450 Da.

Aufgabe 5 Plotten Sie die Exponential- und Pareto-Verteilungen, die sie aus den Peaks mit Masse über 450 Da geschätzt haben, sowie eine Kernel-Density Estimation der tatsächlichen Intensitätsverteilung dieser Peaks. Alle Kurven sollten in einen Plot, damit sie verglichen werden können. Eine Legende sollte die verschiedenen Kurven beschreiben.

Aufgabe 6 Plotten Sie das Spektrum und beschriften Sie die Achsen. Siehe Figure 1. für einen Beispielpplot eines Spektrums. Hinweis: Nutzen Sie die *vlines* Funktion zum Plotten.

