1. Übung Skriptsprachen in der Bioinformatik

Sommersemester 2014

Sascha Winter, Kai Dührkopp

Ausgabe: 11.08.2014

Aufgabe Massenspektrometrie ist ein Verfahren in der (Bio-)chemie zum Analysieren von Molekülen (insbesondere Proteinen). Dabei wird ein Compound/Molekül in kleinere Teilmoleküle fragmentiert und die Massen und Häufigkeiten der Bruchstücke gemessen. Von einem Molekül liegen in der Regel mehrere Messungen vor (jede Messung ist ein Spektrum, eine Liste von (Masse, Intensität) Tupeln).

Das Massbank-Format ist ein Dateiformat in dem solche Massenspektren in einer Datenbank abgelegt werden. Das Format hat leider etliche Schwächen: So werden verschiedene Messungen ein und desselben Moleküls nicht zusammengefasst sondern als voneinander unabhängige Dateien abgespeichert. Peaks werden als Tripple (Masse, Intensität, relative Intensität) abgespeichert. Die relative Intensität kann ignoriert werden. Der Name des Moleküls ist der Substring im Feld *RECORD_TITLE* vor dem ersten Semicolon. Der Molekülname der ersten Datei *CE000001.txt* mit "*RECORD_TITLE: Erythromycin; LC-ESI-ITFT; MS2; CE 35 eV; [M+H]+*" ist beispielsweise *Erythromycin*.

1. Schreibt ein Python Programm, dass aus einem Verzeichnis mit Massbank-Files alle Messungen eines Compounds/Moleküls zusammenfasst. Das Programm soll ein Dictionary zurückgeben, dessen Key der Name des Moleküls ist und dessen Value eine Liste aller (Masse, Intensität) Tuple aller Files dieses Moleküls ist.

   Hinweis: Alle Messungen desselben Moleküls erfolgen immer hintereinander. Wenn ihr die Dateien sortiert vorliegen habt, könnt ihr mit groupby die Dateien nach ihrem Namen gruppieren. Alternativ könnt ihr auch in einer Schleife jede Datei durchgehen und ihre Peaks in ein Dictionary einfügen. Ihr braucht lediglich eine Funktion, die für eine Massbank-Datei deren Compound-Namen und die enthaltenen Peaks zurückgibt.
2. Oft interessieren eher die relativen Intensitäten. Normalisiert daher die Spektren, in dem ihr für jeden Compound die Intensität seiner (Masse, Intensität) Tuple durch die Gesamtsumme der Intensitäten dieses Compounds dividiert.

   Beispiel: [(140.021, 27), (132.001, 44), (116.0102, 18)] wird normalisiert zu [(140.021, 0.3), (132.001, 0.5), (116.0102, 0.2)] in dem alle Intensitäten durch (27+44+18) dividiert werden
3. Eine alternative Form der Normalisierung ist, durch die maximale Intensität statt der Summe zu dividieren. Normalisiert die Spektren noch einmal in dem ihr die maximale Intensität zur Normalisierung heranzieht.

   Beispiel: [(140.021, 27), (132.001, 44), (116.0102, 18)] wird normalisiert zu [(140.021, 0.6), (132.001, 1.0), (116.0102, 0.4)] in dem alle Intensitäten durch 44 dividiert werden
4. Gebt für jeden Compound eine Datei aus deren Name dem Namen des Compounds entspricht. Alle darin enthaltenen Zeilen sind Tuple (Masse, Intensität, Summen-Normalisierung,

Maximum-Normalisierung) die hintereinander weg mit Whitespaces getrennt geschrieben werden.