Übungen zum Sequenzanalyse-Praktikum

Universität Bielefeld, SoSe 2014
Prof. Dr. Jens Stoye · M.Sc. Nina Luhmann · M.Sc. Linda Sundermann
http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2014summer/SequaPrak
praktikum-seqan@CeBiTec.Uni-Bielefeld.DE

Übungsblatt 3 vom 28.04.2014
Abgabe am Donnerstag, den 01.05.2014

Wir wollen eine BLOSUM Matrix selbst berechnen. Zur Berechnung der Matrix benötigen wir lokale Alignments ohne Gaps, wie wir sie in der BLOCKS Datenbank (`http://blocks.fhcrc.org`) finden können. Angenommen wir haben einen Block, der eine konservierte Region einer Proteinfamilie repräsentiert. Aus diesem Block werden Kosten für Matches und Mismatches berechnet. Für jede Spalte des Blocks werden zunächst die Anzahl der Matches und jeder Art von Mismatch für alle Paare von Sequenzen innerhalb des Blocks gezählt.

Beispiel

Wir betrachten eine Spalte aus einem Block, die 9 x `A` und 1 x `S` enthält. Daraus ergeben sich 8+7+...+1=36 mögliche `AA` Paare, 9 `AS` oder `SA` Paare und kein `SS` Paar. Die Häufigkeiten aller beobachteten Paare in jeder Spalte jedes Blocks werden summiert. Wenn also ein Block w Aminosäuren breit ist, und s Sequenzen enthält, trägt dieser Block mit $ws(s-1)/2$ Aminosäure-Paaren bei (in unserem Beispiel (1 x 10 x 9)/2 = 45 Paare). Als Ergebnis dieser Zählung erhalten wir eine Häufigkeitstabelle, die angibt, wie oft jedes der 20+19+...+1=210 verschiedenen Aminosäure-Paare in den Blöcken auftritt. Aus dieser Tabelle berechnen wir eine Log-Odds-Matrix als Verhältnis aus den beobachteten und den zufällig erwarteten Häufigkeiten.

Damit ihr eure (Zwischen)ergebnisse vergleichen könnt, sucht euch die Blocks mit den AC *IPB001303D* und *IPB001525A* aus der BLOCKS Datenbank. Lest einen Block in eure Datei ein und errechnet die BLOSUM Matrix anhand der folgenden Schritte:

Aufgabe 1

f_{ij} sei die Anzahl der beobachteten Häufigkeiten der Aminosäuren i und j ($1 \leq j \leq i \leq 20$).
Schreibe eine Funktion, die für jeden Block der Eingabe die Häufigkeiten der Aminosäure-Paare zählt und diese in einer Matrix f_{ij} speichert.

Aufgabe 2

Die beobachteten Wahrscheinlichkeiten eines Auftretens des Paares i, j ergeben sich aus:

$$q_{ij} = f_{ij} / \sum_{i'=1}^{20} \sum_{j'=1}^{i'} f_{i'j'}$$

In unserem Beispiel mit 9 x `A` und 1 x `S` wäre $f_{AA} = 36$ und $f_{AS} = 9$, $q_{AA} = 36/45 = 0,8$ und $q_{AS} = 9/45 = 0,2$.
Schreibe eine Funktion, die aus der Matrix f_{ij} die Matrix q_{ij} berechnet.

Aufgabe 3

Als nächstes wollen wir die erwarteten Wahrscheinlichkeiten für ein Paar i, j abschätzen. Dazu schauen wir uns wieder unser Beispiel an:
36 Paare haben ein `A` in beiden Positionen, so dass die erwartete Wahrscheinlichkeit für ein `A` in einem Paar [36 + (9/2)] / 45 = 0,9 ist. Für `S` ergibt sich: (9/2) / 45 = 0,1. Allgemein ist die Wahrscheinlichkeit für das Auftreten der Aminosäure i in einem i, j Paar:

$$p_i = q_{ii} + \sum_{j \neq i} q_{ij}/2$$

Schreibe eine Funktion, die das Array p_i aus q_{ij} berechnet.

Aufgabe 4

Die erwartete Wahrscheinlichkeit e_{ij} des Auftretens eines i, j Paares ist $p_i p_j$ für $i = j$ und $p_i p_j + p_j p_i = 2p_i p_j$ für $i \neq j$. In unserem Beispiel ist die erwartete Wahrscheinlichkeit für `AA` 0,9 x 0,9 = 0,81, die von `AS` + `SA` ist 2 x (0,9 x 0,1) = 0,18, und die von `SS` ist 0,1 x 0,1 = 0,01.
Schreibe eine Funktion, die die Matrix e_{ij} aus dem Array p_i berechnet.

Aufgabe 5

Die BLOSUM Matrix ergibt sich nun als das Verhältnis der beobachteten und erwarteten Wahrscheinlichkeiten:

$$s_{ij} = log_2(q_{ij}/e_{ij})$$

Um vergleichbare Scores zu PAM Matrizen zu bekommen, wird dieser Wert noch mit 2 multipliziert und zum nächsten Integer gerundet.
Schreibe eine Funktion, die die BLOSUM Matrix berechnet und ausgibt.

Eine Beispielausgabe zum Vergleich für die oben genannten ACs findet ihr nach dem Praktikum im Wiki.

In eurem Protokoll sollt ihr die einzelnen Schritte zur Erstellung der BLOSUM Matrix kurz erläutern. Euer Programmcode muss ausführbar sein, die Programmausgabe braucht ihr aber nicht ins Protokoll zu übernehmen.