

Übungen zum Sequenzanalyse-Praktikum

Universität Bielefeld, WS 2018/19

Dr. Roland Wittler · M.Sc. Tizian Schulz

<http://gi.cebitec.uni-bielefeld.de/teaching/2018winter/sequaprak>

praktikum-seqan@CeBiTec.Uni-Bielefeld.DE

Übungsblatt 5 vom 06./07.11.2018

Abgabe bis Sonntag bzw. Montag, 24:00 Uhr.

Um eine BLOSUM Matrix zu berechnen benötigen wir lokale Alignments ohne Gaps, wie sie in der BLOCKS Datenbank (<http://blocks.fhcrc.org>) zu finden sind. Aus einem Block werden Scores 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.

Das folgende Beispiel und die Herleitung des Scores sind analog zur Original-Publikation der BLOSUM Matrix (Henikoff, Henikoff: "Amino acid substitution matrices from protein blocks", *PNAS*, 89(22):10915–10919, 1992).

Beispiel

Wir betrachten eine Spalte aus einem Block, die 9 mal A und 1 mal S enthält. Daraus ergeben sich $8 + 7 + \dots + 1 = 36$ 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 $w \cdot s \cdot (s - 1) / 2$ Aminosäure-Paaren bei (in unserem Beispiel $(1 \cdot 10 \cdot 9) / 2 = 45$ Paare). Als Ergebnis dieser Zählung erhalten wir eine Häufigkeitstabelle, die angibt, wie oft jedes der $20 + 19 + \dots + 1 = 210$ 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, sind die Blocks mit den Bezeichnern *IPB001303D* und *IPB001525A* aus der BLOCKS Datenbank im Ordner `/prj/seqan/Praktikum/Blocks/` gegeben. Lest jeweils einen Block ein und errechnet die BLOSUM Matrix anhand der folgenden Schritte:

Aufgabe 1

f_{ij} sei die Anzahl der beobachteten Häufigkeiten eines Aminosäure-Paares i, j oder j, i .

Was passiert, wenn ein Paar (i, j) nicht in den Daten beobachtet wird? (Bitte beantworten.) Um dieses Problem zu umgehen, führen wir *Pseudocounts* ein, d.h. wir addieren bei *jedem* Paar 1 auf, als hätten wir es einmal mehr beobachtet.

Blöcke können auch "Platzhalter" wie X enthalten. Paare, die solche Platzhalter enthalten, sollen nicht gezählt werden.

Schreibe eine Funktion, die für jeden Block der Eingabe die Häufigkeiten der Aminosäure-Paare – inklusive Pseudocounts – zählt und diese in einer Matrix f_{ij} speichert. Dein Programm soll den Dateinamen eines Blocks übergeben bekommen und diesen einlesen und verarbeiten.

Aufgabe 2

Die beobachteten Wahrscheinlichkeiten für das Auftreten einer Substitution 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 mal A und 1 mal S wäre $f_{AA} = 36$ und $f_{AS} = 9$, $q_{AA} = 36/45 = 0,8$ und $q_{AS} = 9/45 = 0,2$. In diesem Beispiel sind Pseudocounts außer Acht gelassen. Die Formel für q_{ij} verwendet jedoch f_{ij} inklusive der Pseudocounts.

Schreibe eine Funktion, die aus der Matrix f_{ij} die Matrix q_{ij} berechnet.

Bitte wenden.

Aufgabe 3

Als nächstes wollen wir die erwarteten Wahrscheinlichkeiten für das Auftreten einer Aminosäure i in einem Paar 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} für das Auftreten eines i, j Paares in unabhängigen Sequenzen 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 $0,9 \cdot 0,9 = 0,81$, die von AS + SA ist $2 \cdot (0,9 \cdot 0,1) = 0,18$, und die von SS ist $0,1 \cdot 0,1 = 0,01$.

Schreibe eine Funktion, die die Matrix e_{ij} aus dem Array p_i berechnet.

Aufgabe 5

Die BLOSUM Matrix ergibt sich nun als Verhältnis der beobachteten Substitutionswahrscheinlichkeiten und der erwarteten Wahrscheinlichkeiten in unabhängigen Sequenzen:

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

In der original BLOSUM62 Matrix werden die Werte in *half-bits* angegeben. Multipliziere dazu die Werte s_{ij} noch mit 2 und runde zum nächsten Integer.

Schreibe eine Funktion, die die BLOSUM Matrix berechnet und in Matrixform ausgibt.

Eine Beispielausgabe zum Vergleich für die oben genannten Bezeichner findet ihr auf der Veranstaltungsseite.

In eurem Protokoll sollt ihr die einzelnen Schritte zur Erstellung der BLOSUM Matrix *kurz* erläutern, d.h., eure Erklärungen sollen zeigen, dass ihr das Vorgehen wirklich verstanden habt. Was bedeuten die Terme in den Formeln? Euer Programmcode soll in *einer* Datei abgegeben werden und muss ausführbar sein. (In Java bitte keine "package"-Anweisung verwenden.) Die Programmausgabe braucht ihr nicht ins Protokoll zu übernehmen.