# Übungen zum Sequenzanalyse-Praktikum

Universität Bielefeld, WS 2019/20
Dr. Roland Wittler · M. Sc. Tizian Schulz
http://gi.cebitec.uni-bielefeld.de/teaching/2019winter/sequaprak
praktikum-seqan@CeBiTec.Uni-Bielefeld.DE

**Übungsblatt 6 vom 26./27.11.2019**
**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 jeder Art von Matches und 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 $8 + 7 + \cdots + 1 = 36$ mögliche `AA` Paare, 9 `AS` oder `SA` Paare und kein `SS` Paar. Die Häufigkeiten der beobachteten Paare werden über alle Spalten summiert. Als Ergebnis dieser Zählung erhalten wir eine Häufigkeitstabelle, die angibt, wie oft jedes der $20 + 19 + \cdots + 1 = 210$ verschiedenen Aminosäure-Paare in den Blöcken auftritt. Aus dieser Tabelle berechnen wir eine Log-Odds-Matrix als Verhältnis aus den beobachteten und den in unabhängigen Sequenzen 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. Hier finden sich Beispielausgaben mit Zwischenergebnissen. Lest jeweils einen Block ein (Übergabe des Dateinamens als Argument) und errechnet die BLOSUM Matrix anhand der folgenden Schritte:

### Aufgabe 1

Sei $f_{ij}$ 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. Um im Folgenden die Beispiele besser nachvollziehen zu können, sind dort in der Regel keine Pseudocounts berücksichtigt. Alle Formeln sollen jedoch unter Berücksichtigung von Pseudocounts implementiert werden.

Blöcke können auch "Platzhalter" wie `X` enthalten. Paare, die solche Platzhalten 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` wären ohne Berücksichtigung von Pseudocounts $f_{AA} = 36$ und $f_{AS} = 9$, $q_{AA} = 36/45 = 0,8$ und $q_{AS} = 9/45 = 0,2$; und mit Berücksichtigung von Pseudocounts $f_{AA} = 36+1$ und $f_{AS} = 9+1$, $q_{AA} = (36+1)/(45+210) \approx 0,145$ und $q_{AS} = (9+1)/(45+210) \approx 0,039$.

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

*Bitte wenden.*

**Aufgabe 3**

Als nächstes schätzen wir die Auftrittswahrscheinlichkeit von Aminosäure $i$ ab, indem wir die relative Häufigkeit von $i$ in einem Aminosäure-Paar berechnen. Dazu schauen wir uns zunächst wieder unser Beispiel an: Ohne Berücksichtigung von Pseudocounts haben 36 Paare ein `A` in beiden Positionen und 9 Paare ein `A` an einer Position, so dass die relative Häufigkeit von `A` in einem Paar $(36 \cdot 2 + 9 \cdot 1)/(45 \cdot 2) = 36/45 + (9/2)/45 = 0,9$ ist. Für `S` ergibt sich: $(9/2)/45 = 0{,}1$.

Allgemein ist die relative Häufigkeit von Aminosäure $i$ in einem 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 Wahrscheinlichkeit $e_{ij}$ für das Auftreten eines $i, j$ Paares in unabhängigen Sequenzen ist $p_i p_j$ für $i = j$ und $p_i p_j + p_j p_i = 2 p_i p_j$ für $i \neq j$. In unserem Beispiel ist die Wahrscheinlichkeit für `AA` (ohne Berücksichtigung von Pseudocounts) $0,9 \cdot 0,9 = 0,81$, die von `AS` oder `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 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 detaillierte Beispielausgabe für einen der oben genannten Bezeichner findet ihr im oben genannten Ordner.

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.