ASE Übung 3 - Log Odds Matrix

Einführung

In dieser Übung soll ein Verständnis für *Log-Odds* Substitutionsmatrizen wie z.B. der Blocks-Substitution-Matrix (BLOSUM) entwickelt werden. Für einen kleinen Datensatz kurzer alinierter Sequenzen mit kleinem Alphabet soll die Matrix von Hand berechnet werden. Dazu sind genaue Anweisungen gegeben.

Eine Log-Odds Matrix M enthält an jeder Position i, j den Wert:

$$M_{ij} = \left\lfloor (2 \times log_2 \frac{q_{ij}}{e_{ij}}) \right\rceil$$

Hierbei ist e_{ij} die erwartete und q_{ij} die beobachtete relative Häufigkeit von Aminosäuren-Substitutionen in den zur Berechnung verwendeten Daten. Diese Werte können als Scores für Substitutionen von Aminosäuren in Sequenzalignmentverfahren verwendet werden.

Für die Berechnung der Matrix soll folgender Beispiel-Datensatz mit reduziertem Alphabet verwendet werden:

```
Block 1: AAIA    Block 2: IWIA
         AALA             IWLA
         AAIL             LWAA
         LALA             LWIA
         AAIW             AWLA
```

Bei den „Blocks“ handelt es sich um kurze, lückenlose, multiple Sequenzalignments stark konservierter Proteinabschnitte. Die folgenden Berechnungen sollen für den gesamten Datensatz und nicht pro Block ausgeführt werden.

Berechnung der Matrix

Für die folgenden Berechnungen ist es zur Übersicht sinnvoll Ergebnisse, welche von zwei Aminosäuren abhängen, jeweils in eine Matrix einzutragen, wobei das verwendete Aminosäuren-Alphabet (`AILW`) die Zeilen und Spaltenindizes bildet.

Nutzen Sie zur Berechnung einen Taschenrechner mit wissenschaftlichem Modus (z.B. kcalc). Sollte Ihr Taschenrechner nicht den Logarithmus zur Basis 2 berechnen können,

berechnen Sie zunächst den Logarithmus zu einer verfügbaren Basis b (meist 10 oder e) und teilen Sie das Ergebnis anschließend durch $\log_b 2$.

1. Berechnen Sie die relative Häufigkeit jeder Aminosäure p_i.
2. Berechnen Sie die erwartete relative Häufigkeit der Aminosäuren-Substitutionen. Für gleiche Aminosäuren handelt es sich um $e_{ii} = p_i^2$ während bei ungleichen Aminosäuren die Vorschrift $e_{ij} = 2p_ip_j$ lautet.
3. Zählen Sie die tatsächlichen Aminosäuren-Substitutionen. Für jedes mögliche Paar von Aminosäuren des Alphabets, müssen die Vorkommnisse des Paares in den Spalten des Datensatzes gezählt werden.
4. Berechnen Sie die relativen Häufigkeiten q_{ij} der Aminosäure-Substitutionen. Hierzu teilen Sie die gezählte Anzahl der Substitutionen pro Aminosäurepaar durch die gesamte Anzahl aller Substitutionen. Pro Spalte gibt es für n Elemente $\frac{n(n-1)}{2}$ Paar-Kombinationen. Für k Spalten der Länge n gibt es entsprechend $k\frac{n(n-1)}{2}$ Substitutionen insgesamt.
5. Berechnen Sie die Log-Odds Matrix-Elemente an den Positionen i, j mit der Formel: $\left\lfloor (2 \times log_2 \frac{q_{ij}}{e_{ij}}) \right\rceil$.

Fragen

1. Wie kommt ein $e_{ij} = 0$ bzw. $q_{ij} = 0$ Wert zustande? Welches Problem tritt dann bei der Berechnung der Matrix auf? Wie kann man dieses Problem vermeiden?
2. Nennen Sie Anforderungen an einen für die Erstellung einer Substitutionsmatrix nach vorgestelltem Verfahren verwendeten Datensatz. Wie viele Proteine sollten verwendet werden? Wie homolog sollten sie sein? Sollten lokale oder globale Alignments verwendet werden? Wie sollten die Alignments aussehen?
3. Wie würden Substitutions-Matrizen aussehen, welche für Alignments unterschiedlich verwandter Proteinpaare verwendet werden können? Nennen Sie einen Ansatz diese verschiedenen Matrizen zu berechnen.