Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, WS 2024
Prof. Dr. Jens Stoye · Leonard Bohnenkämper

https://gi.cebitec.uni-bielefeld.de/teaching/2024winter/sa

Übungsblatt 7 vom 28.11.2024
Abgabe am 05.12.2024 bis 10:00 Uhr (morgens)

Aufgabe 1 (Wiederholung q-Gramm Distanz) (6 Punkte)

1. Zeige, dass die q-Gramm Distanz eine Pseudo-Metrik ist, d.h. zeige Nichtnegativität, Symmetrie und Dreiecksungleichung.
2. Zeige dass die q-Gramm Distanz für $q = 3$ keine Metrik ist.
3. Zeige, dass die q-Gramm Distanz für jedes beliebige q keine Metrik ist.
4. Wir schränken nun die q-Gramm Distanz auf Sequenzen der Länge k ein.
   (a) Zeige, dass für $k = q$ die q-Gramm Distanz eine Metrik ist.
   (b) Welches ist das kleinste $k > q$, sodass die q-Gramm Distanz keine Metrik für Σ^k ist?

Aufgabe 2 (Wiederholung Filter) (6 Punkte)
Kann man die folgenden Distanzen als Filter für die Edit-Distanz mit Einheitskosten nutzen? Finde für Distanz d_a eine Funktion f, sodass garantiert ist, dass $f(d_a(x, y)) \leq d_e(x, y)$, wobei $f(d) > 0$ für $d > 0$. Diskutiere kurz, ob der Filter Sinn ergibt, d.h. für welchen Threshold er einsetzbar ist und welche theoretische Laufzeit benötigt wird.

1. $$d_1(x, y) = \begin{cases} 0 \text{ wenn } x \text{ Subsequenz von } y \\ 1 \text{ sonst} \end{cases}$$
2. $$d_2(x_1 \dots x_n, y_1 \dots y_m) = d_{\text{Hamming}}(x_1 \dots x_k, y_1 \dots y_k)$$
   wobei $k = \min(n, m)$.
3. $d_3(x, y)$ die minimalen Kosten eines Alignments zwischen x und y unter Einheitskosten, aber Gap-Open-Kosten 2 und Gap-Extend-Kosten 1.

Aufgabe 3 (Wiederholung Alignment) (4 Punkte)

1. Berechne ein optimales lokales Alignment zwischen den Strings $x =$ `GAG` und $y =$ `GTAG` mit den Kosten für Match +2, Mismatch −1 und Gap −1.
2. Finde ein optimales globales Alignment zwischen x und y unter dem gleichen Scoreschema.
3. Warum nutzt man für lokale Alignments Scores und nicht Kosten? Welches optimale lokale Alignment existiert unter jedem Kostenmodell, egal welche Strings aligniert werden?
4. Beweise oder Widerlege: Für jedes Kostenmodell gibt es ein Scoreschema, sodass jedes Alignment mit minimalen Kosten, maximalen Score hat und jedes Alignment mit nicht minimalen Kosten, nicht maximalen Score hat.

(Bitte wenden!)

Aufgabe 4 (Σ-Baum) (4 Punkte)

Gegeben sei die Sequenz $s = \texttt{BACA}$.

1. Zeichne den kleinsten Σ-Baum T und den kleinsten kompakten Σ^+-Baum T', die alle Suffixe von s darstellen.
2. Gib für T und T' jeweils die Menge der Worte $x \in \Sigma^*$ an, für die $node(x)$ definiert ist.
3. Welche Menge $words(T)$ von Worten wird durch T dargestellt? Gibt es einen Unterschied zu $words(T')$?