Stefan Kurtz **Universität Hamburg**
Abteilung für Genominformatik **Zentrum für Bioinformatik**

Grundlagen der Sequenzanalyse
Wintersemester 2022/2023
Übungen zur Vorlesung: Ausgabe am 13.12.2022

Aufgabe 8.1 (3 Punkte) Seien u und v zwei Sequenzen der Längen m bzw. n. Sei $lcslen(u,v)$ die Länge der längsten gemeinsamen Subsequenz von u und v.[1] Sei die Kostenfunktion δ für alle Editoperationen $\alpha \to \beta$ wie folgt definiert:

$$\delta(\alpha \to \beta) = \begin{cases} 0 & \text{if } \alpha = \beta \\ 1 & \text{if } \alpha = \varepsilon \text{ or } \beta = \varepsilon \\ \infty & \text{otherwise} \end{cases}$$

Ein Alignment mit einer Ersetzung eines Zeichens durch ein anderes Zeichen hat die Kosten ∞. Ein optimales Alignment von zwei Sequenzen enthält daher keine Ersetzung eines Zeichens durch ein anderes Zeichen.

Beweisen Sie, dass die folgende Gleichung gilt:

$$lcslen(u,v) = \frac{m + n - edist_\delta(u,v)}{2} \tag{1}$$

Beispiel: Sei u = `FREIZEIT` und v = `ZEITGEIST`. Dann ist `EIEIT` die längste gemeinsame Subsequenz von u und v und daher ist $lcslen(u,v) = 5$. Weiterhin gilt $edist_\delta(u,v) = 7$, denn

```
F-REI-Z-EI-T
-Z-EIT-GEIST
```

ist ein optimales Alignment mit Kosten 7. Damit ist

$$\frac{m + n - edist_\delta(u,v)}{2} = \frac{8 + 9 - 7}{2} = \frac{17 - 7}{2} = 5 = lcslen(u,v).$$

Aufgabe 8.2 (4 Punkte) Überlegen Sie sich, wie man den Algorithmus zur Berechnung der Editdistanz von zwei Sequenzen so erweitern kann, dass eine Transposition, d.h. Vertauschung zweier benachbarter Zeichen, kostenlos ist.

Formal bedeutet dies, dass es für ein Alphabet $\mathcal{A}$ für alle $a, b \in \mathcal{A}$ eine weitere Editoperation $ab \to ba$ gibt. Um diese zu bewerten, wird neben der aus dem Vorlesungsskript bekannten Kostenfunktion δ eine Kostenfunktion $\delta' : \mathcal{A}^2 \times \mathcal{A}^2 \to \mathbb{R}$ definiert, die folgende Eigenschaften besitzt:

- $\delta'(ab \to ba) = 0$ für alle $a, b \in \mathcal{A}$
- $\delta'(ab \to cd) > 0$ für alle $a, b, c, d \in \mathcal{A}$ mit $a \neq d$ oder $b \neq c$

[1] Zur Erinnerung: Der Begriff der längsten gemeinsamen Subsequenz wurde in Aufgabe 6.2. definiert.

Beschreiben Sie in einer Datei `permutation_align.tex`, welche zusätzlichen Kanten im Editgraph benötigt werden, um die Transposition zu berücksichtigen. Geben Sie schließlich die Rekurrenzgleichung zur Berechnung der DP-Matrix für das erweiterte Modell an und zeigen Sie, dass die Berechnung der DP-Matrix weiterhin in $O(mn)$ möglich ist.

Punkteverteilung:

- 3 Punkte für die richtige Rekurrenz
- 1 Punkt für den Nachweis der Laufzeit in $O(mn)$

Aufgabe 8.3 (1 Punkt) Im Kontext der Maximal Matches Distanz wurden die Begriffe links-rechts und rechts-links Zerlegung eingeführt. Gegeben seien die Sequenzen $u = agtgcacacatc$ und $v = atcacacttagc$.

1. Bestimmen Sie die links-rechts Zerlegung von u bzgl. v.
2. Bestimmen Sie die rechts-links Zerlegung von u bzgl. v.
3. Bestimmen Sie die links-rechts Zerlegung von v bzgl. u.
4. Bestimmen Sie die rechts-links Zerlegung von v bzgl. u.

Geben Sie diese Werte in der Datei `maximal_matches.tex` an.

Aufgabe 8.4 (4 Punkte) Wir betrachten hier das DNA-Alphabet $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ und Sequenzen, die ausschließlich über diesem Alphabet gebildet werden. Sei $q > 0$. In der Vorlesung wurde gezeigt, wie man alle $w \in \mathcal{A}^q$ als Integer im Wertebereich von 0 bis $4^q - 1$ codieren kann. Die entsprechende Codierung $\overline{w}$ ist wie folgt definiert:

$$\overline{w} = \sum_{i=1}^{q} \overline{w[i]} \cdot r^{q-i}$$

Dabei ist $r = 4$ und die Zeichen des DNA-Alphabet werden entsprechend ihrer Ordnung durch die Zahlen von 0 bis 3 wie folgt codiert: $\overline{\mathtt{A}} = 0$, $\overline{\mathtt{C}} = 1$, $\overline{\mathtt{G}} = 2$, und. $\overline{\mathtt{T}} = 3$.

Integer-Codierungen für aufeinander folgende q-Worte kann man in konstanter Zeit auf Basis der folgenden Gleichung berechnen:

$$\overline{xc} = (\overline{ax} - \overline{a} \cdot r^{q-1}) \cdot r + \overline{c} \tag{2}$$

Diese gilt für alle $a, c \in \mathcal{A}$ und $x \in \mathcal{A}^{q-1}$.

In realen Anwendungen der Integer-Codierungen muss man beide Stränge einer DNA-Sequenz gleichzeitig betrachten, d.h. zusätzlich auch die Integer-Codierung des reversen Komplements eines q-Wortes berechnen.

Sei die Funktion wc : $\mathcal{A} \to \mathcal{A}$ definiert durch $\mathtt{A} \mapsto \mathtt{T}, \mathtt{C} \mapsto \mathtt{G}, \mathtt{G} \mapsto \mathtt{C}, \mathtt{T} \mapsto \mathtt{A}$. Offensichtlich liefert $\mathrm{wc}(b)$ für alle $b \in \mathcal{A}$ das Watson-Crick-Komplement des Nukleotids $b \in \mathcal{A}$. Die Integer-Codierung $\overleftarrow{w}$ des reversen Komplements von w ist wie folgt definiert:

$$\overleftarrow{w} = \sum_{i=1}^{q} \overline{\mathrm{wc}(w[q-i+1])} \cdot r^{q-i} \tag{3}$$

Bearbeiten Sie folgende Teilaufgaben und dokumentieren Sie Ihre Lösung in einer Datei mit dem Namen `intcode_rev.tex`

1. Entwickeln Sie in Analogie zu Gleichung (2) eine Gleichung, um für alle $a, c \in \mathcal{A}$ und $x \in \mathcal{A}^{q-1}$ die Integer-Codierung $\overleftarrow{xc}$ aus $\overleftarrow{ax}$ in konstanter Zeit zu berechnen. Begründen Sie Ihre Lösung anhand einer grafischen Darstellung, in Analogie zu `qgram_slides.pdf` (Frame 12). Falls Sie Ihre Darstellung in tikz erstellen wollen, finden Sie entsprechenden LaTeX-Code dieser Darstellung in der Datei `intcode_fwd.tex`. 2 Pkte
2. Geben Sie in einer Tabelle mit vier Spalten für alle $b \in \mathcal{A}$ die drei Werte $\mathrm{wc}(b)$, $\overline{b}$ und $\overline{\mathrm{wc}(b)}$ an. $\frac{1}{2}$ Pkt
3. Entwickeln Sie aus dieser Tabelle einen möglichst einfachen arithmetischen Ausdruck, durch den man weder unter Verwendung von wc noch mit einer Fallunterscheidung den Wert $\overline{\mathrm{wc}(b)}$ aus $\overline{b}$ berechnen kann. 1.5 Pkt

Bitte die Lösungen zu diesen Aufgaben bis zum 18.12.2022 um 22:00 Uhr an gsa@zbh.uni-hamburg.de schicken.