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

Grundlagen der Sequenzanalyse
Wintersemester 2022/2023
Übungen zur Vorlesung: Ausgabe am 08.11.2022

In Aufgabe 3.2. soll ein Induktionsbeweis geführt werden. Für alle Studierenden, die noch keine oder wenig Erfahrungen mit Induktionsbeweisen haben, werde ich am Ende der Vorlesung am 8.11.2022 einen kurzen Vortrag über dieses Thema halten.

Aufgabe 3.1 (3 Punkte) Gegeben sei eine Sequenz der Länge 1 und eine Sequenz der Länge $n \geq 0$. Zeigen Sie durch vollständige Induktion über n, dass es $2n + 1$ verschiedene Alignments dieser Sequenzen gibt. Formatieren Sie den Beweis in einer Datei mit dem Namen `aligns1n.tex` (falls Sie LaTeX verwenden) oder `aligns1n.txt`, falls Sie Textformat bevorzugen.

Punkteverteilung: 1 Punkt für den richtigen Ansatz. 0.5 Punkte für den Induktionsanfang. 1.5 Punkte für den Induktionsschritt.

Aufgabe 3.2 (4 Punkte) Sei δ eine Kostenfunktion, so dass für alle Editoperationen $\alpha \to \beta$ die folgenden Eigenschaften gelten:

$$\delta(\alpha \to \beta) = \delta(\beta \to \alpha)$$
$$\delta(\alpha \to \beta) = 0 \iff \alpha = \beta$$

Zeigen Sie, dass dann für alle $u, v \in \mathcal{A}^*$ die folgenden Eigenschaften gelten:

1. $edist_\delta(u, v) = edist_\delta(v, u)$
2. $edist_\delta(u, v) = 0 \iff u = v$

Dabei steht das Zeichen $\iff$ für *genau dann wenn*, also die Implikation in beide Richtungen. Sie müssen also in Ihrem Beweis von 2. zunächst $edist_\delta(u, v) = 0$ annehmen und daraus $u = v$ folgern. Dann müssen Sie annehmen, dass $u = v$ gilt und daraus dann $edist_\delta(u, v) = 0$ schliessen.

Der Beweis von 1. kann in ähnlicher Weise als Widerspruchsbeweis geführt werden, wie der Beweis zu

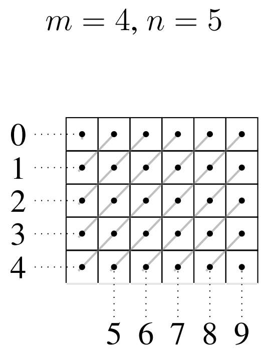
$$edist_\delta(u, v) = edist_\delta(u^{-1}, v^{-1})$$

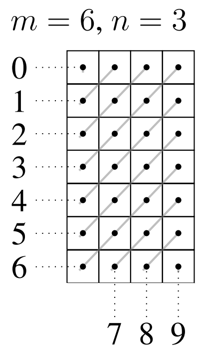
in der GSA-Vorlesung.

Schreiben Sie Ihre Lösung in die Datei `edist_metrix.tex`.

Aufgabe 3.3 (6 Punkte) Sei E_δ die Matrix, die zur Lösung des Editdistanzproblems für zwei Sequenzen u und v der Länge m bzw. n berechnet wird. Anstatt einer spalten- oder zeilenweisen Berechnung der Werte in der $(m + 1) \times (m + 1)$-matrix E_δ können diese auch antidiagonalweise berechnet werden. Eine Antidiagonale umfaßt Werte auf einer Linie von Südwest nach Nordost.

Beispiel. Hier sehen Sie für $(m, n) = (4, 5)$ (links) bzw. für $(m, n) = (6, 3)$ (rechts) eine Darstellung der Antidiagonalen. In der graphischen Darstellung werden für jede Antidiagonale a die Zentren der Matrixeinträge in E_δ, die zu dieser Diagonale gehören, mit einer Gerade verbunden. In der tabellarischen Darstellung sind die Einträge $E_\delta(i, j)$ einer Antidiagonalen durch das Indexpaar (i, j) angegeben.





Antidiagonale	zugehörige Indexpaare
0	(0, 0)
1	(0, 1), (1, 0)
2	(0, 2), (1, 1), (2, 0)
3	(0, 3), (1, 2), (2, 1), (3, 0)
4	(0, 4), (1, 3), (2, 2), (3, 1), (4, 0)
5	(0, 5), (1, 4), (2, 3), (3, 2), (4, 1)
6	(1, 5), (2, 4), (3, 3), (4, 2)
7	(2, 5), (3, 4), (4, 3)
8	(3, 5), (4, 4)
9	(4, 5)

Antidiagonale	zugehörige Indexpaare
0	(0, 0)
1	(0, 1), (1, 0)
2	(0, 2), (1, 1), (2, 0)
3	(0, 3), (1, 2), (2, 1), (3, 0)
4	(1, 3), (2, 2), (3, 1), (4, 0)
5	(2, 3), (3, 2), (4, 1), (5, 0)
6	(3, 3), (4, 2), (5, 1), (6, 0)
7	(4, 3), (5, 2), (6, 1)
8	(5, 3), (6, 2)
9	(6, 3)

Beantworten Sie die folgenden Fragen:

1. Sei *Antidiags*(m, n) die Anzahl der Antidiagonalen in E_δ. Geben Sie eine Gleichung an, anhand derer man *Antidiags*(m, n) berechnen kann.
2. Geben Sie die Eigenschaft aller Paare (i, j) an, die zu der gleichen Antidiagonalen a gehören.
3. Von welchen Antidiagonalen in E_δ hängen die Werte der Antidiagonale $a \geq 2$ ab?

Geben Sie, analog zu Algorithmus 1 (Computation of the edit distance, `scompare_slides.pdf`) Pseudocode für die antidiagonalweise Berechnung von E_δ für zwei beliebige Sequenzen u und v der Längen m und n an. Der Pseudocode muss zunächst den Wert in Antidiagonale 0 berechnen, dann alle Werte in Antidiagonale 1, dann alle Werte in Antidiagonale 2 etc. Hinweise:

- Für jede Antidiagonale a berechnet man zwei Indexwerte $i_{\min}$ und $i_{\max}$, die den minimalen bzw. maximalen Zeilenindex aller Werte auf Antidiagonale a angibt.
- Für ein Indexpaar (i, j) müssen Sie im Pseudocode nicht die komplette Rekurrenz angeben, sondern können schreiben:

  Berechne $E_\delta(i, j)$ entsprechend der Rekurrenz aus der Vorlesung.

Formatieren Sie Ihre Antwort in einer Datei mit dem Namen `EDtab_antidiagonal.tex` (falls Sie LaTeX verwenden) oder `EDtab_antidiagonal.txt`, falls Sie Textformat bevorzugen.

Punktevergabe:

- jeweils 1 Punkt für die richtige Antwort auf die drei Fragen.
- 3 Punkte für korrekten Pseudocode.

Bitte die Lösungen zu diesen Aufgaben bis zum 13.11.2022 um 22:00 Uhr an gsa@zbh.uni-

hamburg.de schicken.