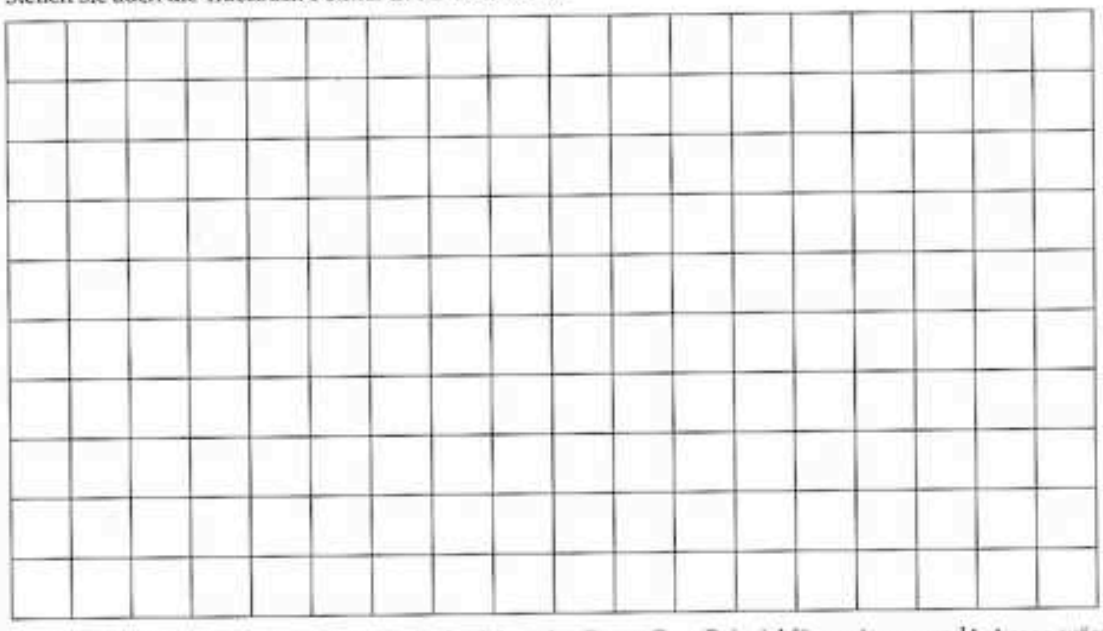
Aufgabe 1 *15 Punkte*

Führen Sie den Algorithmus von Smith-Waterman für die Sequenzen

$$x = \texttt{acbaaac} \quad \text{und} \quad y = \texttt{cbabaab}$$

durch. Verwenden Sie dabei das folgende Scoring Schema: Match=1, Mismatch=−1, Indel=−1, lineare Gap-Kosten.

Stellen Sie auch die Traceback-Pointer in der Matrix dar.



Schreiben Sie das gefundene Alignment in der folgenden Form: Zum Beispiel für caabaacc und babcaaa wäre das Alignment $\begin{array}{c|c|c} \texttt{ca} & \texttt{ab-aa} & \texttt{cc} \\ \hline \texttt{b} & \texttt{abcaa} & \texttt{a} \end{array}$.

Aufgabe 2 *4+4+6+6=20 Punkte*

Wir berechnen ein globales Alignment von zwei Strings x und y, die beide die Länge n haben, mit einer *banded Version des Algorithmus von Needleman-Wunsch.*

Banded bedeutet: es werden in der DP-Matrix und der Traceback-Matrix nur Einträge an solchen Positionen (i, j) gemacht (und in der Rekurrenz berücksichtigt), für die gilt: $|i - j| \leq B$. (Der Parameter B kontrolliert die Breite des Bandes.)

Das Scoring Schema entspricht der Edit-Distanz: Match=1, Mismatch=−1, Indel=−1, lineare Gap-Kosten.

(a) Schätzen Sie die Laufzeit ab in Abhängigkeit von n und B.

(Nehmen Sie einfach als gegeben an, dass der Speicher für $((n + 1) \times (n + 1))$-Matrizen zur Verfügung steht; die Speicherorganisation und -verwaltung soll hier vernachlässig werden.)

(b) Wir berechnen ein banded global Alignment mit $B = 1$ und finden ein globales Alignment mit k Mismatches und/oder Indels. Für welche Werte von k können wir davon ausgehen, dass wir die optimale Lösung gefunden haben? Betrachten Sie $k = 0, 1, 2, 3, 4$ und begründen Sie Ihre Antwort.

(c) Zeigen Sie: Wenn $k < 2B$, dann ist die Lösung optimal. Geben Sie außerdem ein allgemeines Beispiel (d. h., eines für beliebiges B), dass dies für $k \geq 2B$ nicht mehr gelten muss.

(d) Wir starten den banded Needleman-Wunsch zunächst mit $B = 1$, und verdoppeln dann B und berechnen den banded Needleman-Wunsch neu, solange, bis wir gemäß Teil (c) sicher sind, die optimale Lösung gefunden zu haben. Schätzen Sie die Laufzeit dieses Algorithmus ab, in Abhängigkeit von der Stringlänge n und der Edit-Distanz k. Vergleichen Sie Ihre Antwort (kurz, nur 1 Satz) mit der zu (a).

Aufgabe 3 *10+5=15 Punkte*

Für die Sequenzen x und y wurden entsprechend dem FastA-Ansatz zunächst die unten angegebenen diagonal Runs von Substrings festgestellt, die der Übersicht halber mit verschiedenen Buchstaben markiert sind.

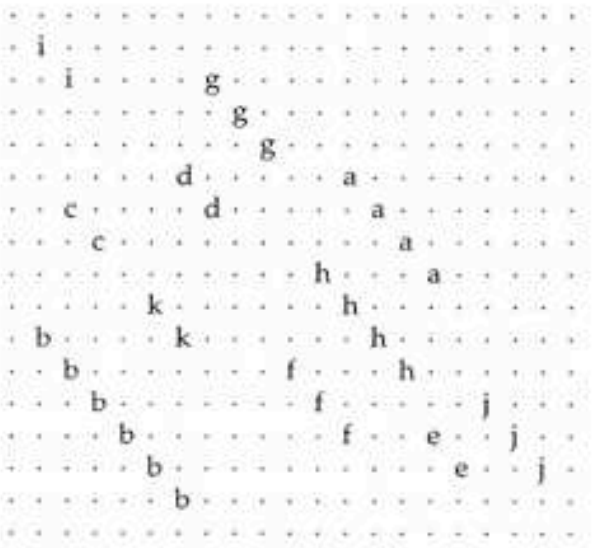
Aus diesen Daten soll nun ein optimales Chaining berechnet werden. Erstellen Sie dazu zunächst einen gerichteten azyklischen gerichteten Graphen mit (positiven) Knoten- und (negativen) Kanten-Scores. Finden Sie dann mit dem in der Vorlesung und Übung behandelten Algorithmus einen gerichteten Pfad mit maximalem Score. Der Score eines Pfades ist die Summe der Scores seiner Knoten und Kanten.

Das Scoring Schema *für das resultierende Alignment* ist: Jede match-Position entsprechend den angegebenen Identitäten zählt 2, alle anderen Positionen des alignments zählen -1 (egal ob es „in Wirklichkeit" vielleicht ein match, mismatch oder indel ist). *Daraus ergeben sich die Knoten- und Kantenscores.*

Beispiele: Wenn man d und h hintereinanderstellt, muss man für die Lücke dazwischen 3 bezahlen („soviel wie Punkte auf dem Weg dazwischen liegen"). Die diagonal Runs f und e sind nicht kombinierbar (da überlappend). Von h nach e kostet es 1 (ein Insert).

(a) Zeichnen Sie den Graphen (mit Knoten- und Kanten-Scores), markieren Sie die Traceback-Kanten und geben Sie den optimalen Pfad an.

(b) Notieren Sie das resultierende Alignment schematisch in der folgenden Form: Beispielsweise der Pfad d, h, e entspräche dem Alignment $\begin{matrix}\texttt{ddMIIhhhh-ee}\\ \texttt{ddM--hhhhIee}\end{matrix}$. (`M` = Mismatch, `I` = Insert, `-` = Space.)



Aufgabe 4 *10+5=15 Punkte*

Gegeben sind zwei Strings x, y und eine Zahl k. (Achtung: k ist Teil der Eingabe.)

(a) Wie kann man in $O(|x| + |y|)$ Zeit feststellen, ob x und y einen gemeinsamen Substring der Länge k besitzen?

(b) Wie kann man in $O(|x| + |y|)$ Zeit einen Substring der Länge k finden, der die meisten diagonal runs der Länge k in einer Dotplot-Ansicht produziert?

(Ob überhaupt ein gemeinsamer Substring der Länge k existiert, kann man gemäß (a) vorab prüfen.)

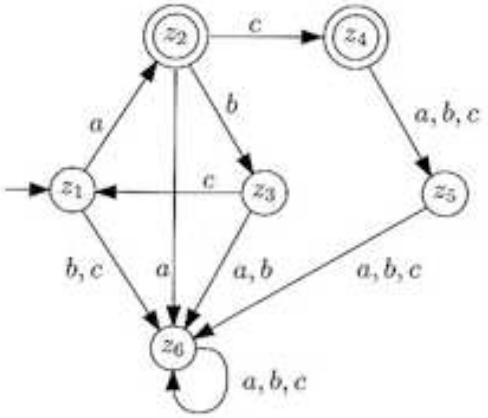
Begründen Sie jeweils auch die behauptete Laufzeit. *(Hinweis: Suffixbaum.)*

Aufgabe 5 *15 Punkte*

Beschreiben Sie die Grundzüge des Blast-Algorithmus für Proteinsequenzen. Gehen Sie dabei auch auf die Rolle der einzelnen Parameter ein, und welche Vor- und Nachteile es hat, wenn man sie größer oder kleiner setzt.

Aufgabe 6 *9+6=15 Punkte*

(a) *Einfach:* Wandeln Sie den folgenden DFA in einen regulären Ausdruck um:



(b) *Schwieriger:* Wir fügen nun noch zusätzlich einen ε-Übergang von z_5 nach z_1 ein. Welche Sprache akzeptiert der resultierende Automat? Geben Sie wiederum einen regulären Ausdruck an.

Aufgabe 7 *2+2+2+2+2=10 Punkte*

Quiz. Welche der folgenden Aussagen sind wahr, welche falsch?

Für jede richtige ja/nein Antwort gibt es 2 Punkte, für jede falsche Antwort gibt es 2 Punkte Abzug. Unbeantwortete Fragen sind neutral. Zusatzregel: Wenn Sie für eine ja/nein Antwort zusätzlich eine korrekte Begründung in ein, zwei Sätzen andeuten, so können Sie damit verhindern, dass für eine eventuelle falsche ja/nein Antwort zu einer anderen Frage Punkte abgezogen werden. Für diese Aufgabe gibt es insgesamt nie weniger als 0 Punkte.

(a) Der Posterior-Decoding-Pfad kann Wahrscheinlichkeit 0 haben, auch wenn der Viterbi-Pfad eine Wahrscheinlichkeit > 0 hat.

(b) Der Viterbi-Pfad hat eine höhere Wahrscheinlichkeit als der Posterior-Decoding-Pfad.

(c) Für jede Position i in der Sequenz gilt immer

$$\sum_q F(q,i)B(q,i) = 1\,.$$

Hierbei läuft q über die Zustände des Hidden Markov Models, und F, B sind die Forward- und Backward-Variablen.

(d) Um die Parameter eines HMMs mit dem Baum-Welch-Algorithmus zu trainieren, benötigt man zu jeder Trainingssequenz die zugehörige Zustandsfolge.

(e) Für die Ereignise A, B, C gelte $A \subseteq B \subseteq C$ (d. h., A impliziert B impliziert C). Dann gilt

$$\Pr(A \mid C) \cdot \Pr(B) \;\leq\; \Pr(A \mid B) \cdot \Pr(C)\,.$$

Aufgabe 8 *15 Punkte*

Gegeben ist die Sprache

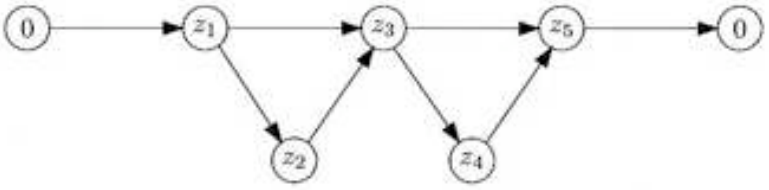
$$L = \left\{ \begin{array}{c} \text{aaaaa}, \\ \text{ababa}, \\ \text{abaaa}, \\ \text{abaa}, \\ \text{aaaa} \end{array} \right\}$$

über dem Alphabet $\Sigma = \{a, b\}$. (Man beachte, dass $aaa, aaba, aaaba \notin L$.) In dem unten stehenden HMM sind bereits Pfeile für alle Übergänge gezeichnet, die eine Wahrscheinlichkeit > 0 haben (dürfen).

Beschriften sie diese Kanten mit Übergangswahrscheinlichkeiten und die Knoten mit Emissionswahrscheinlichkeiten, so dass L gleich der Menge aller Worte ist, die mit einer Wahrscheinlichkeit $\geq 1/10$ generiert werden.

Es sind 8 Übergangswahrscheinlichkeiten und 10 Emissionswahrscheinlichkeiten festzulegen. Verwenden Sie dazu die Zahlen aus der folgenden Liste:

$$0, 0, 0, \quad \frac{1}{5}, \frac{1}{5}, \quad \frac{1}{3}, \frac{1}{3}, \quad \frac{2}{3}, \frac{2}{3}, \quad \frac{4}{5}, \frac{4}{5}, \quad 1, 1, 1, 1, 1, 1, 1.$$



Aufgabe 9 *10 Punkte*

Gegeben ist ein HMM mit den Übergangswahrscheinlichkeiten

$$A := \begin{array}{c|ccc} & 0 & p & q \\ \hline 0 & 0 & 1 & 0 \\ p & a_{p,0} & a_{p,p} & a_{p,q} \\ q & a_{q,0} & a_{q,p} & 0 \end{array}$$

und beliebigen Emissionswahrscheinlichkeiten b für die Zustände p, q und das Alphabet $\{X, Y\}$.

Gesetzt den Fall, das HMM hat die Zeichenkette XYX generiert (also genau diese 3 Zeichen). Was ist dann die bedingte Wahrscheinlichkeit, dass das Y aus dem Zustand p heraus emittiert wurde? Geben Sie das Ergebnis als geschlossene Formel an. Zur Herleitung können Sie ad hoc argumentieren oder den Algorithmus zum Posterior Decoding verwenden.