Übungen zur Vorlesung Sequenzanalyse

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

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

Übungsblatt 4 vom 7.11.2024
Abgabe am 14.11.2024 bis 10:00 Uhr (morgens)

Aufgabe 1 (Modifikation des Alignment-Graphen) (8 Punkte)
Mehrere augenscheinlich sehr verschiedene Alignment-Probleme lassen sich durch den Alignment-Graphen modellieren. Hier betrachten wir nun, wie wir den Graphen an Spezialfälle von Alignmentproblemen anpassen können.
Gegeben seien zwei Sequenzen x, y. Entwirf den Alignment-Graphen, sodass der kürzeste Pfad vom Startknoten v_S zum Endknoten v_E einem optimalen Alignment in folgenden Problemen entspricht. Gib Knoten, Kanten und Gewichte explizit und formal an.
Versuche in jedem Fall den Alignment-Graphen mit den wenigsten Kanten zu finden.

1. Edit-Flip: Unter Einheitskosten, finde ein globales Alignment A mit geringsten Kosten zwischen x und y. Zusätzlich zu den gewöhnlichen Alignment-Spalten sind Flip-Spalten $\binom{ab}{ba}$ für zwei beliebige Zeichen $a, b \in \Sigma, a \neq b$ mit Kosten 1 erlaubt.

2. Wir wollen für zwei proteincodierende DNA-Sequenzen x, y ein globales Alignment unter Einheitskosten finden. Weil wir vermuten, dass dann Gaps von Vielfachen von 3 wahrscheinlicher sind (kein "Frameshift"), wollen wir zusätzlich zu den Edit-Spalten mit Einheitskosten, eine Spalte mit drei aufeinanderfolgenden Gaps mit Kosten 1 erlauben, d.h. $\binom{a\ \ b\ \ c}{-\ \ -\ \ -}$ und $\binom{-\ \ -\ \ -}{d\ \ e\ \ f}$ für alle Kombinationen von $a, b, c, d, e, f \in \Sigma$.

3. Wir nehmen an, x sei eine ungespleißte RNA-Sequenz, d.h. x besteht aus $k+1$ Exons und k Introns: $x = x_0 z_1 x_1 \dots z_k x_k$, wobei $(z_i) \in \Sigma^+$ für $i = 1 \dots k$ und $(x_j) \in \Sigma^+$ für $j = 0 \dots k$. Der Einfachheit halber nehmen wir an, dass weitere Editierungen oder alternatives Spleißen nicht vorkommen, d.h. die vollständig gespleißte Sequenz ist immer $x_0 x_1 \dots x_k$. Zudem gehen wir davon aus, dass die Aufteilung von x in $x_0, z_1, \dots$ bekannt ist. Wir nehmen an, y ist noch nicht vollständig gespleißt, aber verwandt mit x. Gesucht ist also ein globales Alignment zwischen einer Sequenz w aus der Menge der möglichen Spleiß-Zwischenprodukte $\{x_0\} \times \{\epsilon, z_1\} \times \{x_1\} \dots \times \{\epsilon, z_k\} \times \{x_k\}$, dessen Alignment-Kosten zu y minimal sind.

4. Wir nehmen an, y bestehe aus nicht exakten Wiederholungen von x. Gesucht ist ein String w und ein Alignment A zwischen w und y mit minimalen Einheitskosten, wobei w aus einer beliebigen Anzahl an Wiederholungen von x bestehen soll, also $w = xx \dots x$.

Bitte wenden!

Aufgabe 2 (Beziehung Edit-Sequenz–Globales Alignment) (4 Punkte)

Gegeben seien die Sequenz $x = \texttt{AACGGCT}$ und die Edit-Sequenz $E = \mathcal{I}_{\texttt{G}}\mathcal{C}\mathcal{I}_{\texttt{G}}\mathcal{C}\mathcal{C}\mathcal{S}_{\texttt{C}}\mathcal{C}\mathcal{D}\mathcal{C}$.

1. Gib das Alignment A an, welches der Edit-Sequenz E, angewandt auf x entspricht und berechne die Projektion $y = \pi_{\{2\}}(A)$.
2. Wie hoch sind die Einheits-Kosten der Edit-Sequenz E?
3. Betrachte nun die Edit-Scores: $\mathcal{I} = \mathcal{D} = -\frac{3}{2}$, $\mathcal{S} = 0$ und $\mathcal{C} = 1$. Was ist dann der Score von E?

Aufgabe 3 (Optimaler Score) (2 Punkte)

Die Rekurrenz zur Berechnung der Edit-Distanz mit Einheitskosten lautet

$$\text{für } 1 \leq i \leq |x|, 1 \leq j \leq |y| : \; D(i,j) = \min \begin{cases} D(i-1,\, j-1) + \mathbb{1}_{\{x[i] \neq y[j]\}} \\ D(i-1,\, j) + 1 \\ D(i,\, j-1) + 1 \end{cases} \quad \text{mit den Basisfällen} \begin{cases} D(0,0) = 0\,, \\ D(i,\, 0) = i \text{ für } 1 \leq i \leq |x| \text{ und} \\ D(0,\, j) = j \text{ für } 1 \leq j \leq |y| \end{cases}$$

Gib die Rekurrenz und ihre Basisfälle zur Berechnung des optimalen Edit-*Scores* von zwei Sequenzen x und y an.

Aufgabe 4 (Berechnung von Alignments) (6 Punkte)

Gegeben seien die Sequenzen $x = \texttt{AGTACCAGCT}$ und $y = \texttt{GATGCGGACG}$ sowie die folgende Score-Funktion: Match = 3, Mismatch = -2, Indel = -2.

1. Berechne alle optimalen *free-end gap* Alignments von x und y.
2. Berechne alle optimalen lokalen Alignments von x und y.

Mache in deiner Lösung deutlich, wie du zu dem Ergebnis gekommen bist, z.B. indem du die berechneten Alignment-Matrizen angibst. Du kannst dazu die Vorlage auf der nächsten Seite nutzen.

Free end gap:

	ϵ	G	A	T	G	C	G	G	A	C	G
ϵ											
A											
G											
T											
A											
C											
C											
A											
G											
C											
T											

Lokal:

	ϵ	G	A	T	G	C	G	G	A	C	G
ϵ											
A											
G											
T											
A											
C											
C											
A											
G											
C											
T											