Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, SS 2022
Prof. Dr. Jens Stoye · Dr. Marília D. V. Braga

https://gi.cebitec.uni-bielefeld.de/teaching/2022summer/sa

Übungsblatt 11 vom 23.6.2022
Abgabe am 30.6.2022 bis 12:00 Uhr (mittags)

Aufgabe 1 (Suchraum multipler Alignments) (6 Punkte)

Um den Wert $D(i_1, i_2, \ldots, i_k)$ eines Knotens im k-dimensionalen Edit-Graphen mittels des grundlegenden Algorithmus (Abschnitt 11.2.1 im Skript) zu berechnen, muss über eine Anzahl an Werten minimiert werden, die der Anzahl Vorgängerknoten im Graphen entspricht. Die Zahl dieser Vorgängerknoten ist unterschiedlich, je nachdem, wo man sich im Edit-Graphen befindet; es gibt also verschiedene *Typen* von Knoten. Betrachte die folgenden Situationen, mit k Sequenzen der Länge 3:

(i) $k = 2$; (ii) $k = 3$; (iii) $k = 4$.

Für die entsprechenden k-dimensionalen Edit-Graphen sind die folgenden Fragen zu beantworten:

1. Wie viele Knoten hat der k-dimensionale Edit-Graph? (Start- und Endknoten des Graphen können ignoriert werden.)
2. Welche verschiedenen Typen an Knoten gibt es, und wie viele eingehende Kanten hat jeder von ihnen? Wie viele Knoten jeden Typs gibt es?
3. Wie viele Berechnungsschritte werden also insgesamt durchgeführt?

Aufgabe 2 (Carrillo-Lipman-Heuristik) (4 Punkte)

1. Wie viele Carrillo-Lipman-Schranken $U_{x,y}$ müssen berechnet werden, um k Sequenzen zu alignieren? Wie ist die asymptotische Laufzeit für die Berechnung der Schranken?
2. Wie muss eine Menge an Sequenzen beschaffen sein, damit die Carrillo-Lipman-Heuristik gut bzw. schlecht funktioniert?
3. Um den Suchraum mittels Carrillo-Lipman-Schranken einzuschränken, verwendet man eine obere Schranke für die optimalen Alignmentkosten. Durch geschicktes „Mogeln“ kann diese Schranke verfeinert werden.
   (a) Diskutiere Vor- und Nachteile dieses „Mogelns“.
   (b) Überlege dir eine Strategie, wie man günstige Werte für die $\epsilon_{(x,y)}$ findet, um in möglichst kurzer Zeit ein garantiert optimales Alignment zu erhalten.

Aufgabe 3 (Center-Star-Approximation) (5 Punkte)

Gegeben sind die Sequenzen $s_1 =$ `ATACT`, $s_2 =$ `ATCT` und $s_3 =$ `GTGT`. Benutze für deine folgenden Berechnungen Einheitskosten.

1. Berechne die *Center-Sequenz* s_c.
2. Erstelle das multiple Alignment A_c und gib seine Sum-of-Pairs-Kosten an.
3. Was kannst du mithilfe deiner Lösung über die optimalen Alignment-Kosten sagen?
4. Beschreibe in eigenen Worten die Laufzeit- und Speicherplatzkomplexität der Center-Star-Approximation. Unterscheide dabei zuerst die einzelnen Phasen und erkläre dann das Gesamtergebnis.

Aufgabe 4 (Divide-and-Conquer-Alignment) (5 Punkte)

Gegeben sind die Sequenzen $s_1 = CTC$, $s_2 = AC$ und $s_3 = AT$. Benutze für deine Berechnungen Einheitskosten.

1. Was ist der Unterschied zwischen einem optimalen und einem C-optimalen Schnitt?
2. Erstelle die Zusatzkostenmatrizen für die drei Sequenzen.
3. Gib alle C-optimalen Schnitte an.
4. Nenne alle möglichen optimalen multiplen Alignments, die dir die C-optimalen Schnitte angeben. Gib auch ihre Kosten an.
5. Wieso verkleinert sich der Suchraum von $\mathcal{O}(n^k 2^k)$ beim Sum-of-Pairs-optimalen multiplen Alignment auf $\mathcal{O}(n^{k-1})$ beim Divide-and-Conquer-Alignment?