Technische Universität München
Fakultät für Informatik
Lehrstuhl für Effiziente Algorithmen
Dr. Volker Heun

Wintersemester 2001/2002
Übungsblatt 11
16. Januar 2002

Algorithmische Bioinformatik

Abgabetermin: Donnerstag, den 24. Januar (15^{00} Uhr) in der Übung

Aufgabe 1

Betrachte folgende Sequenzen $s_1 = ACGTGC$, $s_2 = AGCC$, $s_3 = ACCTG$ und $s_4 = AGGCTT$. Der optimale Abstand für die paarweise Sequenzen-Alignments ist rechts angegeben. Hierbei gilt $w(x,y) = 1$ und $w(x,x) = 0$ für alle $x \neq y \in \overline{\Sigma}$. Konstruiere für diese Sequenzen ein mehrfaches Sequenzen-Alignment mit Hilfe der Center-Star-Methode.

d	s_1	s_2	s_3	s_4
s_1	0	3	2	4
s_2	3	0	3	3
s_3	2	3	0	3
s_4	4	3	3	0

Hinweis: *Die als freiwillig gekennzeichneten Aufgaben werden nicht bewertet und sind für die Klausurzulassung ohne Bedeutung.*

Freiwillige Aufgabe 1

Zeige, dass es für jedes $n \in \mathbb{N}$ eine Menge von Sequenzen $S = \{s_1, \ldots, s_k\}$ mit $|s_i| \geq n$ gibt, so dass es einen Sequenz $s \in S$ gibt, die als Zentrum bei der Center-Star-Methode eine nicht-konstante Approximationsgüte liefert. Hierbei gilt $w(x,y) = 1$ und $w(x,x) = 0$ für alle $x \neq y \in \overline{\Sigma}$.

Freiwillige Aufgabe 2

Betrachte die Sequenzen $S = \{s_1, \ldots, s_k\}$. Sei M_c das mehrfache Sequenzen-Alignment (MSA), das mit der Center-Star-Methode erzeugt wurde, wobei $c \in [1:k]$ der Index einer Sequenz ist, die als Zentrum gewählt wurde. Es bezeichne $D(s_i, s_j)$ den Wert des durch M_c induzierten paarweisen Alignments von s_i und s_j und $D(M_c)$ den Wert des MSA M_c. Dann gilt $d(s_i, s_j) \leq D(s_i, s_j)$, $d(s_c, s_j) = D(s_c, s_j)$ und $D(M_c) = \sum_{i=1}^{k} \sum_{j=i+1}^{k} D(s_i, s_j)$. Weiter sei M^* ein optimales MSA für S und $D^*(s_i, s_j)$ der Wert des durch M^* induzierten paarweisen Alignments von s_i und s_j. Dann gilt $D(M^*) = \sum_{i=1}^{k} \sum_{j=i+1}^{k} D^*(s_i, s_j)$. Aus der Vorlesung kennen wir folgendes

Lemma *Es existieren mehr als $\left\lfloor \frac{k}{r} \right\rfloor$ Sterne S_i mit s_i als Zentrum und $M(i) \leq \frac{2r-1}{r-1} \cdot M$, wobei $M(i) := \sum_{j=1}^{k} d(s_i, s_j)$ und $M := \min\{M(i) \mid i \in [1:k]\}$.*

O.B.d.A. gilt für die Nummerierung der Sequenzen in S, dass $M(1) \leq \cdots \leq M(k)$. Wir wählen jetzt $c := \left\lfloor \frac{k}{r} \right\rfloor + 1$ und wählen $\varepsilon := \frac{M(c)}{M}$. Dann gilt: $M(c) = \varepsilon \cdot M$ und $\varepsilon \in [1, \frac{2r-1}{r-1}]$.

Zeige, dass $\dfrac{D(M_c)}{D(M^*)} \leq 2 + \dfrac{1}{r-1}$ ist.

Erinnerung:

Die Klausur findet am **4. Februar 2002** um **17:45 Uhr** im **Großen Physikhörsaal (1200)** statt!