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

Wintersemester 2001/2002
Übungsblatt 9
19. Dezember 2001

Algorithmische Bioinformatik

Abgabetermin: Donnerstag, den 10. Januar (15[00] Uhr) in der Übung

Aufgabe 1

Sei $s = s_1 \cdots s_n \in \Sigma^*$ eine Zeichenfolge. Eine Zeichenfolge $s' = s_1 \cdots s_\ell \in \Sigma^*$ heißt *Teilfolge* von s, wenn es eine streng aufsteigende Folge $(k_1, \ldots, k_\ell) \in \mathbb{N}^\ell$ gibt (d.h., es gilt $k_1 < k_1 < \cdots < k_\ell$), so dass $s'_i = s_{k_i}$ gilt.

Beispiel: eure ist eine Teilfolge von *Sequenzenvergleich*.

Seien $s = s_1 \cdots s_n \in \Sigma^*$ und $t = t_1 \cdots t_m \in \Sigma^*$ zwei Zeichenfolgen. Konstruiere einen Algorithmus, der eine längste Zeichenfolge $u = u_1 \cdots u_\ell$ findet, die sowohl eine Teilfolge von s als auch von t ist.

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

Freiwillige Aufgabe 1

Sei $z = z_1 \cdots z_n \in \Sigma^*$ eine Zeichenreihe. Eine Wiederholung von z ist ein Alignment $(\bar{s}, \bar{t})$ von zwei Teilwörtern $(s, t) = (s_i \cdots s_j, t_k \cdots t_\ell)$ von z mit $i < k$ und $j < \ell$, wobei es jedoch nicht erlaubt ist, das Zeichen an Position i mit sich selbst zu alignieren.

Eine optimale Wiederholung von z ist eine Wiederholung von z mit maximalem Score bezüglich des verwendeten Ähnlichkeitsmaßes. Hierbei gilt für die Kostenfunktion w des Ähnlichkeitsmaßes: $w(a, -) + w(-, b) \geq w(a, b)$ für alle $a, b \in \Sigma$.

Beispiel: Sei $s = s_1 \cdots s_{10} = ababababab$ eine Zeichenfolge. Dann ist

1. $\begin{pmatrix} a & b & a & b & a & b & a & - \\ - & b & a & b & a & b & a & b \end{pmatrix}$ für $(s_1 \cdots s_7, s_2 \cdots s_8)$ keine Wiederholung;
2. $\begin{pmatrix} a & b & a & b & a & b & a & b & - & - & - & - & a & b \\ a & b & - & - & - & - & a & b & a & b & a & b & a & b \end{pmatrix}$ für $(s_1 \cdots s_8, s_3 \cdots s_{10})$ eine Wiederholung;
3. $\begin{pmatrix} a & b & a & b & a & b & a & b \\ a & b & a & b & a & b & a & b \end{pmatrix}$ für $(s_1 \cdots s_8, s_3 \cdots s_{10})$ eine optimale Wiederholung.

Zeige zuerst, dass das Alignment einer optimalen Wiederholung die Diagonale der Dynamic Programming nicht berührt (d.h., dass das Beispiel 2 kein optimale Alignment sein kann).

Gib dann eine auf lokalem Sequence Alignment basierende Methode an, um eine optimale Wiederholung innerhalb einer gegebenen Zeichenreihe $s \in \Sigma^*$ zu finden.

Fröhliche Weihnachten und einen guten Rutsch ins neue Jahr!