Technische Universität München
Fakultät für Informatik
Lehrstuhl für Effiziente Algorithmen
Dr. Hanjo Täubig
Johannes Krugel

Sommersemester 2009
Übungsblatt 8
19. Juni 2009

Algorithmische Bioinformatik I

Aufgabe 1

Für $x, y \in \Sigma^*$ wird eine Sequenz $z \in \Sigma^n$ als gemeinsame Supersequenz (shortest common supersequence) von x und y bezeichnet, wenn zwei streng monoton wachsende Folgen $(i_1, \ldots, i_{|x|}) \in [1 : n]^{|x|}$ und $(j_1, \ldots, j_{|y|}) \in [1 : n]^{|y|}$ gibt, so dass $z_{i_1} \ldots z_{i_{|x|}} = x$ und $z_{i_1} \ldots z_{j_{|y|}} = y$.
Konstruieren Sie einen möglichst effizienten Algorithmus, der für $x, y \in \Sigma^*$ eine kürzeste gemeinsame Supersequenz bestimmt. Analysieren Sie die Laufzeit.
Hinweis: Sie können dabei auf die Funktion zur Berechnung der längsten gemeinsamen Teilsequenz (longest common subsequence) zurückgreifen.

Aufgabe 2

Sei $w : \overline{\Sigma}_0 \to \mathbb{R}^+$ eine Kostenfunktion für ein Distanzmaß d, wobei $w(a, -) = w(-, a) = \gamma$ für alle $a \in \Sigma$ für ein $\gamma \in \mathbb{R}^+$. Sei weiter $\rho \in \mathbb{R}^+$ und seien $s \in \Sigma^n$ sowie $t \in \Sigma^m$. Konstruieren Sie einen möglichst effizienten Algorithmus, der ein optimales Alignment von s und t liefert, sofern $d(s, t) \leq \rho$. Man beachte, dass eine Laufzeit von $O(nm)$ nicht für alle ρ und γ effizient ist.

Aufgabe 3

Sei $s = s_1 \cdots s_n$ eine Zeichenreihe der Länge n. Ein k-mer t[1] von s ist eine Teilwort der Länge k von s, d.h. es existiert ein $i \in [1 : n - k + 1]$ mit $s_i \cdots s_{i+k-1} = t$. Für eine Zeichenreihe s sei $\mathcal{M}_k(s) = \{s_i \cdots s_{i+k-1} : i \in [1 : n - k + 1]\}$ die Multimenge der k-mere von s.
Zeigen Sie, dass es zwei Zeichenreihen s und s' gibt, so dass $\mathcal{M}_4(s) = \mathcal{M}_4(s')$ ist. Dies bedeutet, dass man aus einer Menge der k-mere die ursprüngliche Zeichenreihe nicht mehr eindeutig rekonstruieren kann.
Wie sieht es für beliebiges k aus?

[1] Auch k-Gramm genannt (beziehungsweise N-Gramm, q-Gramm, ...).