Übungen zur Vorlesung Sequenzanalyse I

Universität Bielefeld, WS 2011/2012
Dr. Alexander Sczyrba · Nina Luhmann · Linda Sundermann

http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2011winter/SequenzAnalyse

Übungsblatt 10 vom 20.01.2012
Abgabe in einer Woche vor Beginn der Vorlesung.

Aufgabe 1 Datenbanksuche (4 Punkte)

1. Welche unterschiedlichen Annahmen werden bei

   a) der Suche nach einem optimalen Alignment von zwei Sequenzen und bei

   b) der Suche nach einer ähnlichen Sequenz in einer Datenbank

   getroffen? Welche Anforderungen werden dadurch an die Software gestellt, die diese Probleme lösen soll?

2. Erkläre die Unterschiede zwischen *on-line* und *index-basierter* Datenbanksuche. Wie sind die Laufzeiten zur Vorverarbeitung und bei der Suche? Welche Laufzeit ergibt sich bei k Suchen?
   In welchen Fällen lohnt sich eine index-basierte Suche, wann ist die on-line Variante vorzuziehen?

Aufgabe 2 Suboptimale Alignments (5 Punkte)

1. Beschreibe in eigenen Worten, was man unter *überlappenden* Alignments versteht.

2. Warum ist man bei der Bestimmung von suboptimalen Alignments in erster Linie an *nichtüberlappenden* Alignments interessiert? Welche Probleme möchte man vermeiden?

3. Beschreibe kurz die Funktionsweise des Waterman-Eggert-Algorithmus.
   Durch welchen Trick kann man in der Praxis die Laufzeit verkürzen?

4. Gegeben sei $x =$ `GTAA`, $y =$ `GCTA`, berechne ein lokales Alignment mit dem Smith-Waterman-Algorithmus und das erste nicht überlappende, suboptimale Alignment nach *Waterman-Eggert*. Verwende dazu Scores mit $\text{score}(\mathcal{C}) = 3$, $\text{score}(\mathcal{S}_{a,c}) = 1$ für $a \neq c$ und $\text{score}(\mathcal{I}_c) = \text{score}(\mathcal{D}_c) = -1$.

Aufgabe 3 Approximatives Stringmatching (4 Punkte)

1. Finde die Endpositionen aller Vorkommen des Patterns $x =$ `ATAT` im Text $y =$ `GGTGATATGTAAAC` mit maximal $k = 1$ Fehlern. Verwende die Cutoff-Variante von *Sellers'* Algorithmus mit Einheits**kosten** und markiere die *last essential indices*.

2. Gib zu jeder gefundenen Endposition alle zugehörigen Alignments an.

3. Warum macht es Sinn, in der Praxis nicht *alle* gefundenen Endpositionen auszugeben? (Stichwort: *Runs*.) Welche Endpositionen würde man im obigen Beispiel nicht ausgeben?