Universität Bielefeld
Technische Fakultät

AG Genominformatik
Prof. Dr. Jens Stoye
Dipl.-Math. Julia Mixtacki
Dipl.-Inform. Arne Hauenschild
Dipl.-Inform. Wiebke Timm

Foundations of Sequence Analysis

Winter 2005/2006

Exercises

Übung 1, Besprechung am 31.10., 01.11 bzw. 03.11.2004.

1. **Hamming Distanz.**

   - Berechnen Sie die Hamming Distanz für alle Paare der folgenden Sequenzen:
     $\mathbf{s}_1$ = `ACGATACTAG`, $\mathbf{s}_2$ = `AGGTCATTGA`, $\mathbf{s}_3$ = `AGGCATTGA`, $\mathbf{s}_4$ = `CGATACTAGA`.
   - Implementieren Sie eine Funktion `hammingDistance` zur Berechnung der Hamming Distanz in **Java**, **Haskell** oder **C**.

2. **Euklidische Distanz und Block Distanz.**

   - Berechnen Sie jeweils die Euklidische Distanz und die Block Distanz für alle Paare der folgenden Vektoren über $\mathbf{Z}^7$:
     $\mathbf{v}_1$ = (`1, 4, 3, 9, 1, 2, 5`),
     $\mathbf{v}_2$ = (`6, 4, 11, -9, -4, 8, 7`),
     $\mathbf{v}_3$ = (`5, 1, 4, 3, 9, 1, 2`).
   - Implementieren Sie die Funktionen `euclidianDistance` und `blockDistance` zur Berechnung der Euklidischen Distanz bzw. der Block-Distanz in **Java**, **Haskell** oder **C**.

3. **Metrik und Hamming Distanz.**

   Seien $\mathbf{u} = u_1u_2 \dots u_n$ und $\mathbf{v} = v_1v_2 \dots v_n$ Strings der Länge $n \in \mathbf{N}$ über dem Alphabet $\mathcal{A}$ ($\mathbf{u}, \mathbf{v} \in \mathcal{A}^n$). Dann ist die Hamming Distanz $h : \mathcal{A}^n \times \mathcal{A}^n \longrightarrow \mathbf{R}$ zwischen $\mathbf{u}$ und $\mathbf{v}$ definiert als

   $$h(\mathbf{u}, \mathbf{v}) = \sum_{\substack{i=1 \\ u_i \neq v_i}}^{n} 1.$$

   - Machen Sie sich klar, dass diese Definition äquivalent ist zu der in der Vorlesung angegebenen Definition.
   - Beweisen Sie, dass die Hamming Distanz h eine Metrik auf dem Raum der Sequenzen $\mathcal{A}^n$ der Länge $n \in \mathbf{N}$ ist.

4. **Anzahl von Teilsequenzen.**

   Seien $\mathbf{u} = u_1u_2 \dots u_n$ und $\mathbf{v} = v_1v_2 \dots v_m$ Strings der Länge $n = |\mathbf{u}|$ bzw. $m = |\mathbf{v}|$.

   Eine *Teilsequenz* von $\mathbf{u}$ und $\mathbf{v}$ ist eine Sequenz von Index-Paaren

   $$(i_1, j_1), \dots, (i_r, j_r),$$

so dass

$$1 \leq i_1 < \ldots < i_r \leq n \quad \textit{und}$$
$$1 \leq j_1 < \ldots < j_r \leq m.$$

Eine Teilsequenz steht für ein Alignment von **u** und **v**. Dabei steht das Index-Paar (i_h, j_h) für die Ersetzung $u_{i_h} \rightarrow v_{j_h}$. Alle Buchstaben von **u** und **v**, deren Index nicht in der Teilsequenz erscheint, gelten als Deletion in **u** bzw. **v**.

(a) Zeigen Sie, dass sich die Anzahl der Teilsequenzen `Subseqs(m,n)` von zwei Sequenzen der Länge m und n folgendermaßen berechnet:

$$\texttt{Subseqs(}m\texttt{,}n\texttt{)} = \sum_{r=0}^{min(m,n)} \binom{m}{r} \cdot \binom{n}{r}.$$

(b) Implementieren Sie die Funktionen `subseqs(m,n)` zur Berechnung der Anzahl der Teilsequenzen.