Chair for Clinical Bioinformatics
Professor Dr. Andreas Keller
Dr. Christina Backes
Dipl.-Ing. Thomas Großmann
B. Sc. Yvonne Saara Gladbach

# *Bioinformatik 1 - Übungsblatt 5*

**Abgabe bis Dienstag 02.12.2014 um 11:59**

Vorname, Nachname:
Matrikelnummer:

**Anmerkungen** :
Die Übungen dürfen in Gruppen bestehend aus max. 2 Personen abgegeben werden. Der Quellcode ist als Quellcodepaket abzugeben, muss auf den Cip-Pool Rechnern kompilieren, gut dokumentiert und getestet sein (Testfälle sind mit abzugeben!). Schicken Sie Ihre Abgabe als tar.gz Paket per Mail an Ihre Tutorin Yvonne Gladbach (yvonne.gladbach@ccb.uni-saarland.de) mit Betreff: "Übung X, Name Y, Name Z". Das Paket darf keine temporären oder binären Dateien enthalten. Beim Entpacken des Paketes soll ein übergeordnetes Verzeichnis erstellt werden, welches den Quellcode enthält. Der Quellcode muss durch ein mitgeliefertes Makefile kompilierbar sein.

## Aufgabe 1 (15 P.):

Um Multiple Sequenz Alignments (kurz MSA) zu bewerten, kann man die so genannte Konsensussequenz verwenden. Gegeben ein MSA A bestehend aus $k$ Sequenzen $A_1, \ldots, A_k$ der Länge $l$. Die Konsensussequenz ist die Sequenz, die an jeder Position $i \in \{1, \ldots, l\}$den Buchstaben mit der größten Häufigkeit in $A$ an Position $i$ hat. Die Distanz von $A$ zu $C_A$ ist gegeben als :

$$\begin{aligned} D(C_A, A) =& D(C_A, (A_1, \ldots, A_k)) \\ =& \sum_{j=1}^{l} (k - count(j, C_A[j])) . \end{aligned}$$

Dabei ist $C_A[j]$ der j-te Buchstabe der Konsensussequenz und $count(j, C_A[j])$ ist gegeben als:

$$count(j, C_A[j]) = |\{i \in [1, k] \,|A_i[j] = C_A[j]\}| .$$

(a) Geben Sie die Konsensussequenz zum folgenden MSA an:

| | | | | |
|---|---|---|---|---|
| A | G | G | - | T |
| A | G | G | C | T |
| A | G | - | C | T |
| A | G | G | G | T |
| A | G | G | G | C |

(b) Berechnen Sie die oben definierte Distanz für das MSA und die zugehörige Konsensussequenz.

(c) Beweisen oder widerlegen Sie folgende Aussage: Seien $C_{A1}$ und $C_{A2}$ zwei Konsensussequenzen zu dem selben optimalen MSA $A$. Dann gilt:

$$D\left(C_{A1}, A\right) = D\left(C_{A2}, A\right)$$

**Aufgabe 2 (10 P.):**
Konstruieren (zeichnen) Sie einen Deterministischen Finiten Automaten (DFA), der aus einem gegebenen Text alle Vorkommen der Sequenzen (und nur dieser Sequenzen) AGGT, ACT, AGC und GCG erkennt. Das zu verwendende Alphabet ist $\Sigma = \{A, C, T, G\}$.

**Aufgabe 3 (15 P.):**
In der Vorlesung wurde die Burrows-Wheeler Transformation vorgestellt. Berechnen Sie die Burrows-Wheeler Transformation und den Index für folgende Sequenz: $MISSISSIPPI\$$. Was ist der Index der Burrows-Wheeler Transformation? Geben Sie auch die Zwischenschritte der Burrows-Wheeler Transformation an.

**Aufgabe 4 (10 P.):**
Geben Sie eine Formel zu Berechnung der Komplexität des folgenden Problems an: Sie haben $k - mere$ und $a$ Fragmente, dazu werden PCR-Produkte der Länge $l$ benutzt und man möchte eine Coverage $c$ haben. Begründen Sie ihre Antwort.

**Zusatzaufgabe (15 P.):**
Implementieren Sie die Erweiterung des Assembly-Problems von Blatt 3 Aufgabe 1 indem Sie nun auch die Scoring-Funktion berücksichtigen für möglich Fehler im Overlap der Reads. Dazu laden Sie sich bitte die geänderte Version der $fragments.fta$ von der Vorlesungsseite herunter. Vergleichen Sie ihr Ergebnis mit dem vom Blatt 3 Aufgabe 1 und begründen Sie die Unterschiede.