Klausur zum Modul

Algorithmen in der Bioinformatik

Sommersemester 2019
12.07.2019

Name: .....

Matrikelnummer: .....

Studiengang: .....

Geben Sie den Lösungsweg immer mit an!

Nur mit blauem oder schwarzem Kugelschreiber schreiben.

Schreiben Sie auf jeden Zettel Ihre Matrikelnummer.

Geben Sie für jede (Teil-)Aufgabe nur eine einzige Lösung ab. Bei mehreren, alternativen Lösungen zu einer Aufgabe wird die Schlechteste bewertet.

Teilnahme an der Klausur erfolgt unter Vorbehalt einer vorhandenen Zulassung.

Aufgabe Nr.:	Punktzahl:	Davon erreicht:
1	14	
2	11	
3	12	
4	13	
$\sum$	50	

Es sind keinerlei Hilfsmittel erlaubt. Bitte schreiben Sie deutlich mit einem schwarzen oder blauen Stift.

1. Sequenzalignments [14]

Finden Sie *alle* besten (maximalen) lokalen Sequenzalignments der Sequenzen `ATACTGGG` und `TGACTGAG` mit dem Smith-Waterman-Algorithmus. Ein Match zählt hierbei $+1$, ein Mismatch -1 und eine Lücke (Gap) -2. Geben Sie die gesamte DP-Matrix mit den für die Alignments relevanten Backtracking-Zeigern an und erläutern Sie jeden Ihrer Schritte.

2. Genomassembly

Gegeben sind die folgenden vier Reads: `TACAGT`, `CAGTC`, `AGTCAG` und `TCAGA`.

a) Wie viele 3-mere sind in diesen Reads (inklusive Duplikate)? Wie viele unterschiedliche 3-mere? Wie viele unterschiedliche 2-mere? [3]

b) Was ist die maximale Anzahl unterschiedlicher 2-mere in einer Menge von n Reads der Länge 100 bei einem Alphabet $\{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$? [3]

c) Zeichnen Sie den De-Bruijn-Graphen, in dem die Kanten 4-meren und die Knoten 3-meren entsprechen. Hat dieser Graph einen Eulerweg? Wenn ja, welcher Sequenz entspricht er? Wenn nein, wieso nicht? [5]

3. Suffix-Bäume

Ein *generalisierter Suffix-Baum* ist ein Suffix-Baum für mehrere Strings $s_1, \ldots, s_k$. Hierbei wird für jeden String s_i, $1 \leq i \leq k$, ein eigenes Terminierungszeichen $\$_i$ verwendet, außerdem wird in den Blättern zusätzlich zur Position notiert, aus welchem String das Suffix stammt. Sie können voraussetzen, dass sich ein generalisierter Suffix-Baum in Zeit $O(\sum_{i=1}^{k} |s_i|)$ aufbauen lässt.

a) Zeichnen Sie einen generalisierten Suffix-Baum für die Strings $s_1 =$ `ALAAF` und $s_2 =$ `HELAU`. [4]

b) Beschreiben Sie einen Algorithmus, der für zwei gegebene Strings s_1 und s_2 über einem Alphabet konstanter Größe den längsten gemeinsamen Substring in Zeit $O(|s_1| + |s_2|)$ ausgibt. [8]

4. Clustering

Gegeben ist folgende Instanz eines k-Means-Clusteringproblems für $k = 2$. Punkte sind zu clusternde Datenpunkte, Kreuze sind die initial gewählten Clusterzentren von Lloyds Algorithmus.

a) Welche Clusterzentren gibt Lloyds Algorithmus bei dieser Instanz zurück? Wie viele Schritte benötigt Lloyds Algorithmus dafür? [3]

b) Geben Sie nun Pseudocode für einen exakten Algorithmus für das k-Means-Clusteringproblem mit $k = 2$ an. Der Algorithmus soll auf vollständiger Enumeration aller Partitionen in zwei nichtleere Cluster basieren. [5]

c) Analysieren Sie nun die Laufzeit Ihres Algorithmus. Beweisen Sie dazu folgende Aussage, zum Beispiel mit Hilfe von vollständiger Induktion: Es gibt $2^{n-1} - 1$ verschiedene Partitionen einer Menge von n Elementen in zwei nichtleere Cluster. [5]