Prof. Dr. Knut Reinert
Kathrin Trappe
René Rahn
Kathleen Gallo
Thomas Krannich

Institut für Informatik
AG Algorithmische Bioinformatik

Algorithmen und Datenstrukturen in der Bioinformatik
Zweite Praktikumsaufgabe WS 12/13

Abgabe Montag, 07.01., 15:00 Uhr

Read-Mapping mit QUASAR

Implementieren Sie ein Programm zum schnellen Finden von semi-globalen Alignments zwischen genomischen Reads und einer genomischen Sequenz T. Dazu soll ihr Programm folgende Schritte ausführen:

a) Einlesen der Sequenz T und der Reads aus Dateien (als Vorlage gegeben)

b) Erzeugen eines q-gram Index über T

c) Für jeden Read

  i) Zählen der überlappenden q-gramme des Reads in Blöcken von T

  ii) Sammeln der Blöcke in T, die mit dem Read mindestens t q-gramme gemein haben

d) Verifizieren der gesammelten Block-Read-Paare, indem der jeweilige Read semi-global gegen den Block aligniert wird.

e) Ausgeben der Textendpositionen von Alignments mit höchstens k Edit-Fehlern in eine Ausgabedatei. Pro Block-Read-Paar soll nur **ein** bestes Alignment ausgegeben werden.

Hinweise:

- Das zugrundeliegende Alphabet ist $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.
- Reads sind kurze Sequenzen der Länge 36–400. Hier haben alle Reads dieselbe Länge.
- Setzen Sie $w = \textit{Readlänge}$, $b = 2w$ und $q = 10$ in QUASAR. S entspricht einem Read und D der genomischen Sequenz.

- Ein semi-globales Alignment aligniert einen ganzen Read gegen einen Teil von T. Passen Sie den Needleman-Wunsch Algorithmus entsprechend an.
- Als erstes Argument erhält ihr Programm den Dateinamen der genomischen Sequenz, als zweites den Dateinamen der Reads, als drittes die erlaubte Fehlerzahl k und als viertes den Dateinamen der Ausgabedatei
- Die genomische Sequenz steht in einer Zeile in der Datei, die Datei mit den Reads enthält pro Zeile genau einen Read. In der Ausgabedatei stehen Zeilen in der Form *ReadSequenz,TextEndPosition,Fehlerzahl*.
- Die Code-Vorlagen finden Sie unter `https://svn.imp.fu-berlin.de/aldabi/WS12/material/aufgabe5`.

Beispiel:

- Datei mit genomischer Sequenz:

  ```
  AAGATTACATTTTTAAAAAAAAAAACAATTT
  ```

- Datei mit Reads:

  ```
  GATACA
  CATTTT
  ```

- Ausgabedatei:

  ```
  GATACA,8,1
  CATTTT,13,0
  CATTTT,29,1
  ```

Bewertung:

- Diese Aufgabe wird mit der doppelten Punktzahl bewertet.
- Die Gruppe mit der schnellsten Lösung (auf einem Rechner mit 8 Kernen) erhält Ruhm und Ehre und einen Riesen-Schoko-Weihnachtsmann. Beachten Sie die Praktikumshinweise unter `https://www.mi.fu-berlin.de/w/ABI/AlDaBiWS12`.