Aufgabe 3: Erste Versuche im Indexieren des Templates

Ulf Leser

Wissensmanagement in der Bioinformatik

q-Gram Index

- Ein q-Gram Index für einen String T ist ein invertiertes File über allen q-Grammen von T
- Beispiel: AGGTAGATGATA, q=2
  - AG: 1,5
  - GA: 6,9
  - GG: 2
  - ...
- Q-Gram Indexe sind linear groß
  - An jeder Position in T beginnt ein q-Gram
  - Also ist die Menge aller Startpositionen im File O(T)
- Berechnung: Scanne T und baue den Index im Hauptspeicher
  - Es gibt höchstens $O(|\Sigma|^q)$ q-Gramme im Index

Suche mit q-Gram Indexen

- Gegeben ein Pattern P und einen q-Gram Index für T. Dann können wir alle Vorkommen von P in T wie folgt finden
  - Berechnen alle q-Gramme von P
  - Finde alle Matches von q-Grammen in P mit q-Grammen in T
  - Existiert für ein q-Gram im Pattern kein Match: Sofort stoppen
  - Prüfe, ob q-Gram Matches in der richtigen Reihenfolge und im richtigen Abstand vorkommen
- Beispiel
  - T=TCGTGTC, q=2, P=GTC
  - TC: 1,6; CG:2; GT: 3,5; TG: 4
  - Matches mit GT: 3,5; Matches mit TC:1,6
  - Matches an Position 5 und 6 haben den richtigen Abstand

1. Stringsuche mit q-Gram Indexen (12 Punkte)

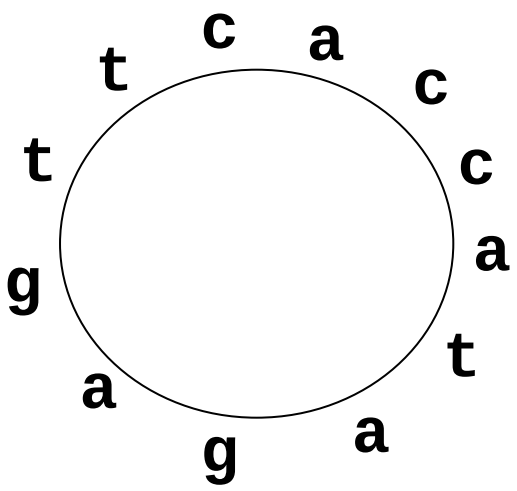
- Implementieren Sie einen Algorithmus, der, gegeben ein Template, einen q-Gram Index berechnet
  - Der Index soll in einer Datei auf Platte geschrieben werden
- Implementieren Sie exakte Stringsuche mit dem q-Gram Index
- Zum Test: Benutzen der Daten wie bisher (Template und Pattern aus Aufgabe 1)
- Ausgabe
  - Pro Pattern
    - Patternlänge
    - Anzahl Fundstellen
    - Startpositionen der ersten 10 Fundstellen
    - Laufzeit
  - Gesamtlaufzeit über alle Pattern

Viele, viele, viele Ideen

- Müssen wir nach allen q-Grammen in P suchen?
  - Nein – eine abdeckende Menge reicht
- Finden aller matchenden q-Gramme?
  - Einmal laden
  - Sortieren und BIN-SEARCH - oder Hashing?
- Welches q?
  - Tja ...
  - Pattern werden zwischen 6 und 50 Zeichen lang sein
- Nach welchen q-Grammen zuerst suchen?
  - Wenn man zwei q-Gramme von P keinmal im richtigen Abstand findet, kann man gleich aufhören
  - Kleine Zwischenergebnisse sind gut
- ...

Zirkuläre Strings und Rotationen

- Ein String a ist eine zirkuläre Rotation eines Strings b, wenn beide Strings gleich lang sind und a aus einem (eventuell auch leeren) Suffix von b konkateniert mit dem verbleibenden Präfix von b zusammengesetzt ist
  - Beispiel: abcdef, defabc, fabcde sind zirkuläre Rotationen von abcdef
- Ein zirkulärer String ist ein „zum Kreis gebogener“ String.
  - Beispiel



2. Aufgabe (8 Punkte)

- (3 Punkte) Geben Sie einen Algorithmus an, der in lineare Zeit entscheidet, ob ein String a eine zirkuläre Rotation eines Strings b ist
  - Unter Rückgriff auf einen beliebigen linearen Stringmatching-algorithmus
- (5 Punkte) Geben Sie einen Algorithmus an, der in lineare Zeit entscheidet, ob ein String a Substring eines zirkulären Strings b ist
  - Achtung: Es kann gelten: |a|>|b|
  - Linear heißt: O(|a|+|b|)
  - Beispiel: aagaaga ist ein Substring von aga

Wettbewerb (0 Punkte)

- Wir vergeben Punkte für zwei Aufgaben
  - Schnellste Indexierung
  - Schnellste Suche
- Beide werden getrennt voneinander gemessen
- Wir verwenden jeweils andere Templates / Pattern
- Speicherverbrauch oder Größe der Indexdatei ist egal
  - Solange es auf gruenau2 laeuft

Programmaufruf

- Das Indexprogram muss wie folgt aufrufbar sein
  - `java –Xmx xyz –jar Assignment3.jar index file1 file2`
  - File1: Dateiname der Templates
  - File2: Dateiname der Indexdatei
  - Ausgabe erfolgt also in file2
  - Schicken Sie uns Ihr xyz
- Das Suchprogramm muss wie folgt aufrufbar sein
  - `java –jar Assignment3.jar search file1 file2`
  - File1: Dateiname der Indexdatei
  - File2: Dateiname der Patterndatei
  - Ausgabe auf STDOUT

Abgabe

- Bis Sonntag, 29.11. bzw. Dienstag, 1.12., 23.59
- Nur per Mail als TXT Datei

1. Lösung der Aufgabe 2 als Text
2. Die Ausgabe der Suche auf den Testdaten
3. Code im Quelltext
4. Ausführbare JAR Datei