Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, WS 2012/2013
Prof. Dr. Jens Stoye · Nina Luhmann · Linda Sundermann

http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2012winter/SequenzAnalyse

Übungsblatt 8 vom 21.12.2012
Abgabe in einer Woche vor Beginn der Vorlesung.

Aufgabe 1 (Burrows-Wheeler Transformation) (5 Punkte)
Gegeben sei der String $t = s\$ = TGATGCATCATGCAT\$$.

1. Berechne die Burrows-Wheeler Transformierte $bwt(t)$ für t.
2. Schreibe $rle(bwt(t))$ als komprimierten String mit Hilfe von *run-length encoding* auf. Fasse dabei nur Buchstaben zusammen, die min. 3 mal hintereinander vorkommen.
3. Überprüfe, ob das Muster $p = ATGC$ im Text t vorkommt. Beschreibe den String Matching Algorithmus unter Verwendung der $bwt(t)$ beispielhaft an der Suche von p.

Aufgabe 2 (Forward-Backward Technik) (6 Punkte)
Gegeben seien die beiden Strings $x = TGCAT$ und $y = TAT$. Die Indizierung beginnt bei 1.

1. Berechne die Matrizen $D, D^{rev} = D^{-1}$ und C unter dem Einheitskostenmodell.
2. Lies alle optimalen Alignments aus C ab und gib diese an.
3. Welche minimalen zusätzlichen Kosten hat ein Alignment unter der Einschränkung, dass $x[2]$ und $y[1]$ aligniert sind?
4. Gib alle Alignments an, die im Vergleich zum optimalen Alignment zusätzliche Kosten von 1 haben.

Aufgabe 3 (Paarweises Alignment mit linearem Speicherbedarf) (4 Punkte)
Gegeben seien zwei Strings $s = GCGAG$ und $t = GATATCG$. Berechne das globale Alignment von s und t unter Verwendung der Divide-and-Conquer Technik und Einheitskosten. Berechne das Teilalignment „normal“– also ohne Divide-and-Conquer Technik –, wenn einer der beiden Strings nur noch Länge eins hat. Simuliere die notwendigen Schritte auf Kopien der Edit-Matrix:

1. Kennzeichne jeweils $m' = \lceil m/2 \rceil$.
2. Gib auch die Backpointer-Matrix M an.
3. Markiere in jedem Rekursionsschritt, welche Teile der Edit-Matrix erneut berechnet werden müssen.

Aufgabe 4 (Divide-and-Conquer (D&C) Technik) (3 Punkte)
Mittels der D&C Technik lassen sich paarweise Alignments mit linearem Speicherbedarf berechnen. In der Vorlesung wurde ein Algorithmus vorgestellt, mit welchem das optimale globale Alignment bestimmt werden kann. Welche Probleme treten auf, wenn die D&C Technik zur Berechnung folgender Alignment-Varianten angewendet wird, und wie lassen sie sich lösen?

1. *Globales Alignment mit affinen Gapkosten (Gotoh)*:
   Bei affinen Gapkosten wird zwischen Kosten für zum Öffnen (*gap-open*) und Erweitern (*gap-extension*) von Gaps unterschieden.
2. *Lokales Alignment (Smith-Waterman)*:
   Lokales Alignment ist das beste globale Alignment aller Substrings zweier Sequenzen.
3. *Ausgabe von k (sub-) optimalen Alignments (Waterman-Eggert)*:
   In einem Alignment wird Backtracing zum Auffinden eines optimalen Alignments angewandt. Häufig gibt es jedoch mehrere optimale Alignments, und in manchen Fällen ist es auch interessant, die k besten – möglicherweise suboptimalen – nicht-überlappenden Alignments zu kennen.

Bitte wenden!

Freiwilliger Weihnachtszettel

* mit Zusatzpunkten *

Aufgabe 5 (MUMs) (****)

Gegeben seien die Strings $s = AGTCGTAGCAGT$ und $t = TCGATTCGCAGAT$. Berechne alle MUMs der minimalen Länge drei mit Hilfe eines generalisierten Suffixbaums und gib ihre Positionen an. Beschreibe dein Vorgehen.

Aufgabe 6 (Free-end-gap-Alignment) (****)

Berechne das free-end-gap-Alignment zwischen den beiden Strings $s = ZUCKERWATTE$ und $t = ERWARTEN$. Verwende folgende Scores: INDEL: -1, SUBSTITUTE: -1 und COPY: $+1$. Gib alle optimalen Alignments wie auch deren Scores an.

Aufgabe 7 (q-gram-Distanz) (****)

Gegeben seien der String $s = ACATGTCGTACATTGTCAT$ und $q = 4$. Zeige anhand des de-bruijn-Graphen für s, ob es weitere Strings s' gibt, deren q-gram-Distanz zu s gleich null ist. Gib einen String an oder begründe, warum es keinen gibt.