Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, WS 2015/2016
Prof. Dr. Jens Stoye · M.Sc. Linda Sundermann

http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2015winter/SequenzAnalyse

Übungsblatt 8 vom 22.12.2015
Abgabe Abgabe am Donnerstag, den 07.01.2016

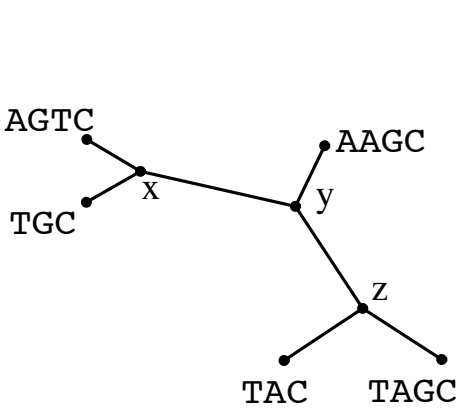
Wir wünschen euch frohe Weihnachten und einen guten Rutsch ins neue Jahr!

Aufgabe 1 (Wiederholung) (3 Punkte)
Welche Themen würdest du gerne für die Klausur wiederholen? Schreibe drei Themen oder konkrete Fragen auf, die dir noch nicht ganz klar sind.

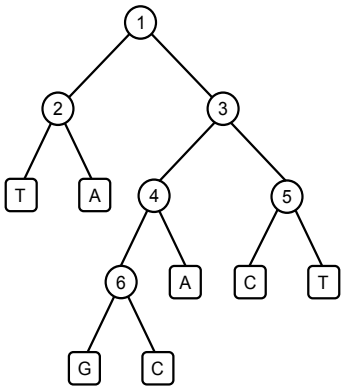
Aufgabe 2 (Center-Star-Approximation) (5 Punkte)
Gegeben sind die Sequenzen $s_1 = \texttt{ATGCT}$, $s_2 = \texttt{ATCT}$ und $s_3 = \texttt{GTGT}$. Benutze für deine folgenden Berechnungen Einheitskosten.

1. Berechne die *Center-Sequenz* s_c.
2. Erstelle das multiple Alignment A_c und gib seine Sum-of-Pairs-Kosten an.
3. Was kannst du über die Optimalität des gefundenen Alignments sagen?
4. Beschreibe in eigenen Worten die Laufzeit- und Speicherkomplexität der Center-Star-Approximation. Unterscheide dabei zuerst die einzelnen Phasen und erkläre dann das Gesamtergebnis.

Aufgabe 3 (Baumalignment) (6 Punkte)
Gegeben seien die folgenden Bäume:



(a)



(b)

1. Um die Kosten für ein Baumalignment exakt zu berechnen, kann der Sankoff-Algorithmus verwendet werden. Erkläre in 3–5 Sätzen, wie dieser Algorithmus funktioniert. Erläutere auch, welche Rolle der Fitch-Algorithmus dabei spielt.
2. Beschrifte die inneren Knoten x, y und z des Baumes in Abbildung 1(a) so, dass die Kosten des Baumalignments möglichst gering sind. (Hier soll kein Algorithmus angewendet werden, sondern "gut geraten" werden.) Berechne die Kosten des entsprechenden Alignments.
3. Berechne für den phylogenetischen Baum in Abbildung 1(b) die sparsamste Beschriftung der inneren Knoten mit Hilfe des Fitch-Algorithmus. Gib dabei für die inneren Knoten jeweils die mit ihnen assoziierten Informationen der Bottom-Up- und der Top-Down-Phase an.

Bitte wenden.

Aufgabe 4 (MUMs) (4 Punkte)

1. Wofür können *Maximal Unique Matches* verwendet werden?
2. Gib einen Algorithmus an, der die MUMs der Länge l oder größer findet. In welcher Komplexitätsklasse liegt der Algorithmus und warum?

Ab hier könnt ihr Extrapunkte sammeln.

Aufgabe 5 (Gotoh) (3* Punkte)

Berechne ein optimales globales Alignment mit affinen Gapkosten von den Sequenzen $x =$ `CGCAT` und $y =$ `CGT` effizient mit Hilfe des Gotoh-Algorithmus und gib dessen Gesamtscore an. Verwende dabei: Score für Match = 2, Kosten für Mismatch = 1, Kosten für Gap-open $d = 1$, sowie Kosten für Gap-extension $e = 0.5$.

Aufgabe 6 (Algorithmen-Tabelle) (9* Punkte)

Mache dir klar, welche Laufzeiten die folgenden Algorithmen haben:
Berechnung der q-gram-Distanz; Berechnung der Maximal-Matches-Distanz; Needleman-Wunsch-Algorithmus; Smith-Waterman-Algorithmus; Algorithmus für Alignment mit free end gaps; Gotoh-Algorithmus; Waterman-Eggert-Algorithmus; Sellers' Algorithmus; Hirschberg-Algorithmus; WOTD-Algorithmus; Algorithmus, um einen String exakt in einem Suffixbaum zu finden; Algorithmus, um mit Hilfe eines Suffixbaums den kürzesten eindeutigen String zu finden; Algorithmus, um maximale Repeats mit einem Suffixbaum zu finden; Konstruktion des Suffixarrays *pos*, wenn der Suffixbaum gegeben ist; exakte Berechnung eines multiplen Alignments; Center-Star-Approximation; Divide-and-Conquer-Alignment-Algorithmus; Sankoffs Algorithmus; Fitch-Algorithmus.
Fülle dazu eine Tabelle in der unten stehenden Form aus. Nenne den Namen des Algorithmus und beschreibe kurz, was er macht, wenn man es nicht schon am Namen erkennen kann. Schreibe die asymptotische Laufzeit auf und erkläre die Bedeutung der Parameter. Verstehe, wie die Laufzeit zu Stande kommt, aber schreibe es nicht auf.

Algorithmus	Zweck	Laufzeit	Parametererklärung
Berechnung der q-gram-Distanz	–	$\mathcal{O}(\sigma^q + m + n)$	σ: Alphabetgröße, q: Länge des q-grams, m und n: Länge der Sequenzen
...			
Needleman-Wunsch-Algorithmus	berechnet ein globales Alignment zwischen zwei Sequenzen	$\mathcal{O}(n \cdot m)$	m und n: Länge der Sequenzen
...			

Aufgabe 7 (Alignment-Typen) (6* Punkte)

Gegeben seien die zwei Sequenzen $x =$ `AATGCT` und $y =$ `TGT`, sowie ein Matchscore von 3 und Mismatch- und Indelkosten von 2. Berechne die folgenden Alignments von x und y in je einer Matrix, wobei du am Ende auch den optimalen Score und ein mögliches optimales Alignment nennen sollst.

1. globales Alignment
2. lokales Alignment
3. semi-globales Alignment (approximative Textsuche)

Mache dir die Unterschiede zwischen den drei Alignment-Typen klar. Wie unterscheiden sich die Initialisierung, die Maximierung im Rekursionsschritt und die Bestimmung des endgültigen Scores?