Klausur zur Vorlesung Sequenzanalyse II

Universität Bielefeld, SoSe 2007
Dr. Sven Rahmann · Dipl.-Inf. Peter Husemann · Dipl.-Biol. Constantin Bannert

http://gi.cebitec.uni-bielefeld.de/teaching/2007summer/sequenzanalyse/

Klausur vom 29.3.2007, 8:30–10:00 Uhr

Name:

Matrikelnummer:

Zulassungsnachweise
☐ Anmeldung
☐ A&D-Prüfung
☐ Übungspunkte

Hinweise: Bitte erst alle Aufgaben einmal durchlesen und mit den einfachsten anfangen. Es sind keine Hilfsmittel (wie z.B. Aufzeichnungen, Taschenrechner, Handys, etc.) zugelassen. Weitere leere Blätter sind bei den Tutoren erhältlich; bitte keine eigenen Blätter verwenden.

Aufgabe	Punkte	max.
1.1		5
1.2		3
1.3		4
2.1		4
2.2		1
2.3		3
2.4		3
3.1		8
3.2		7
3.3		8
4.1		8
4.2		6
4.3		2
5.1		4
5.2		4
5.3		2
Summe		72

Aufgabe 1 Metriken:

1. Gib die allgemeine Definition einer Metrik an.
2. Sei Σ ein endliches Alphabet. Was bedeuten jeweils Σ^n, Σ^*, Σ^+?
3. Ist die q-gram Distanz $d_q : \Sigma^* \times \Sigma^* \to \mathbb{N}$ im allgemeinen eine Metrik auf Σ^*? Wenn ja, gib eine Begründung; wenn nein, zeige an einem Beispiel, welche Eigenschaft einer Metrik verletzt ist.

Aufgabe 2 Maximal-Matches-Distanz und Edit-Distanz:

1. Gegeben sei ein String $s \in \Sigma^n$ und $t := s\#s\#s$ mit $\# \notin \Sigma$. Berechne die Maximal-Matches-Distanzen $\delta(s\|t)$ und $\delta(t\|s)$.
2. Welche Schranke ergibt sich daraus für die Standard-Edit-Distanz $d(s,t)$?
3. Berechne $d(s,t)$ exakt (Hinweis: Wie lang ist t?).
4. Welchen (globalen) Edit-Score erhält man zwischen s und t, wenn man das Scoring-Schema $+1$ für einen match, 0 für einen mismatch und $-1/2$ für einen indel verwendet? Wie ist der Zusammenhang zur Edit-Distanz?

Aufgabe 3 Globales Alignment mit "free end gaps":

1. Wie sieht der Alignment-Graph $G = (V, E)$ aus? Gib genaue Definitionen der Knotenmenge V und der Kantenmenge E an. Zeichne den Graphen auch schematisch.
2. Gib explizit die zugehörigen Rekursionsformeln (mit Initialisierung und Abschluss) im score-maximierenden Fall an. Jede verwendete Variable muss erklärt werden.
3. Wende den Algorithmus auf $s =$ ABBAB und $t =$ ABAAA an, wobei der indel-Score -2 (aber beachte free end gaps!), der mismatch-Score -1 und der match-score $+1$ betragen soll. Gib den optimalen score und ein optimales Alignment an.

Aufgabe 4 Suffixbaum und -array:

1. Zeichne den Suffixbaum von $s =$ bcabcabc\$. Annotiere die Blätter mit den Suffix-Startpositionen und bilde so das Suffixarray `pos`. Beachte dabei die Ordnung \$ < a < b < c.
2. Bestimme das `lcp`-Array. Wie kann man es aus dem Baum ablesen?
3. Wie kann man in Linarzeit das "inverse" Suffixarray `rank` aus `pos` berechnen (Pseudocode)?

Aufgabe 5 Komplexität von Alignmentverfahren:

1. Wir wollen zwei bakterielle Genome der Größe jeweils 6 MB mit Hilfe des Smith-Waterman-Algorithmus vergleichen. Unser PC schafft es, eine Million Edit-Matrix-Einträge pro Sekunde zu berechnen. Pro Matrix-Eintrag spreichern wir einen Score-Wert (`int`, 4 Bytes) und einen backpointer (1 Byte). Wie viel Speicherplatz würden wir benötigen? Wie lange würde der Vergleich in Stunden dauern?
2. Solche Vergleiche werden offenbar aber täglich durchgeführt. Erkläre kurz(!) zwei Methoden, mit denen man das in der Praxis macht, und wie das Laufzeit- und Speicher-Problem dabei vermieden wird (keine Details, nur die wesentliche Idee).
3. Globales oder lokales exaktes Alignment von zwei Sequenzen der Längen m und n mit *linearen* Gapkosten kostet $O(mn)$ Zeit. Mit *allgemeinen* Gapkosten sogar $O((m+n)mn)$. Erläutere, mit welchem Trick man die Zeitkomplexität bei *affinen* Gapkosten wiederum auf $O(mn)$ reduzieren kann.