Prof. Dr. Knut Reinert
Sandro Andreotti
FB Mathematik und Informatik

Freie Universität Berlin

Fortgeschrittene Algorithmen in der Bioinformatik (P4): Sequence and Structure Analysis Abschlussklausur SS 08

Name: Matrikelnummer:

Viel Erfolg!

A1	A2	A3	A4	A5	A6	A7	$\sum$
10	12	16	13	16	12	11	90

Aufgabe 1: Horspool, Wu-Manber — 3 + 3 + 4 = 10 Punkte

Beschreiben Sie den Algorithmus von Wu-Manber zum multiple string matching. Gehen Sie dabei auf folgende Punkte ein:

(a) Was ist, wie benutzt man, und wozu dient die SHIFT table?

(b) Was ist, wie benutzt man, und wozu dient die (erste) HASH table?

(c) Was ist der Vorteil des Wu-Manber Algorithmus gegenüber dem trivialen Erweiterung des Horspool Algorithmus für multiple strings?

Aufgabe 2: PEX, Pidgeonhole Principle — 3 + 3 + 6 = 12 Punkte

Ein Text T der Länge 1000 enthält ein Vorkommnis Occ eines Patterns P der Länge 50 mit höchstens 6 Fehlern. In wieviele Stücke muss man P zerlegen (also $P = P_1P_2 \dots P_k$), um sicherzustellen, dass ...

(a) T eines der Stücke exakt (ohne Fehler) enthält?

(b) T eines der Stücke mit höchstens drei Fehlern enthält?

(c) Beweisen sie das Lemma:
Sei Occ ein approximatives Vorkommen eines strings in einem pattern P mit k Fehlern. Sei weiter $P = p^1, \dots, p^j$ die Konkatenation von Teilen von P sowie $a_1, \dots, a_j$ nichtnegative, ganze Zahlen so dass $A = \sum_{i=1}^{j} a_i$. Dann existiert ein $i \in 1, \dots, j$, so dass es einen substring in Occ gibt der p^i mit $\lfloor a_i k/A \rfloor$ Fehlern matched.

Aufgabe 3: Komparative RNA-Analyse — 3 + 3 + 3 + 7 = 16 Punkte

(a) Was ist eine stochastische, kontextfreie Grammatik (SCFG)?

(b) Was ist die Chomsky-Normal Form?

(c) Gegeben eine SCFG, was berechnet der Inside-Algorithmus?

(d) Geben sie den inside-Algorithmus in pseudocode an.

Aufgabe 4: Quasar **5 + 3 + 5 = 13 Punkte**

(a) Formulieren sie und beweisen sie das q-gram Lemma für ungapped shape?

(b) Gilt das q-gram Lemma auch für gapped shapes?

(c) Falls ja, ist das Lemma dann auch scharf? Begründen sie ihre Antwort mit einem Beispiel?

(d) Falls nein, geben sie ein Gegenbeispiel an.

Aufgabe 5: Suffix arrays **3 + 3 + 4 + 6 = 16 Punkte**

(a) Definieren sie ein suffix array für einen string und die lcp Tabelle.

(b) Geben sie an, wie man in einem Suffix array mit Hilfe der lcp Tabelle effizient sucht.

(c) Wie ist die Laufzeit für die Suche eines Patterns der Länge m in einem string der Länge n?

(d) Geben sie den Algorithmus von Kasai zur Linearzeitberechnung der lcp Tabelle an.

Aufgabe 6: Chaining **5 + 7 = 12 Punkte**

(a) Welche Operationen unterstützt die beim Chaining-Algorithmus verwendete Datenstruktur? Beschreiben Sie jeweils kurz, was sie bewirken bzw. was ihr Ergebnis ist.

(b) Beschreiben Sie den Chaining-Algorithmus für L_1-Gap-Kosten. *Alternativ (halbe Punktzahl)*: Beschreiben Sie den Chaining Algorithmus ohne Gap-Kosten.

Aufgabe 7: Multiple Match refinement **3 + 8 = 11 Punkte**

(a) Welches Problem adressiert man beim Multiple Match Refinement?

(b) Gegeben sei die untige Menge von Segment Matches $\{S_1, S_2, S_3\}$. Geben Sie das eine minimale Menge von refined segment matches grafisch an, wobei sie passende Projektionen wählen dürfen. Begründen sie ihre Lösung kurz.