Sommersemester 2006 Abgabe am 09. Juni 2006

Prof. Dr. Stefan Posch, Dipl.-Bioinform. André Gohr, Dipl.-Bioinform. Jan Grau

9. Übung „Algorithmen der Bioinformatik I“

1. Gegeben sei ein Alphabet A und das Metazeichen $\star$ mit $\star \notin A$. Es gelte die Konvention, daß das Metazeichen $\star$ einen beliebigen String aus A^* repräsentiert. Nun sei das Alphabet $\bar{A} = A \cup \{\star\}$ gegeben. Jeder String aus $\bar{A}$ repräsentiert somit eine Menge von Strings aus A^*, wir bezeichnen sie als Suchmuster. Wir wollen im folgenden solche Muster in einem String T mit Hilfe des Suffixbaums $ST(T)$ suchen.

   a) Gegeben sei das Suchmuster $S_1 \star S_2 \in \bar{A}$ mit $S_1, S_2 \in A^*$. Konstruieren Sie einen Algorithmus, der das Musters einmal in T findet (bzw. das Nichtauftreten feststellt). (3 Punkte)

   b) Gegeben sei nun das allgemeine Suchmuster $S_1 \star S_2 \cdots \star S_k$. Konstruieren Sie einen Algorithmus, der das Musters einmal in T findet (bzw. das Nichtauftreten feststellt). (2 Punkte)
   Schätzen Sie die Laufzeit ihres Algorithmus ab. (2 Punkte)

2. Skizzieren Sie den generalisierten Suffixbaum für die folgende Menge von Strings:

   $\{$ ababbababba, abbaabbaabba, babbababbab $\}$ (3 Punkte)

3. Welche Schritte sind zum Entfernen eines Strings S_i aus einem generalisierten Suffixbaum notwendig? Der Suffixbaum liege dabei wie in der Vorlesung angegeben mit einem gemeinsamen Terminationssymbol für alle Strings sowie mit Listen für die Markierungen in den Blättern vor. (3 Punkte)
   Läßt sich die Aufgabenstellung in $\mathcal{O}(|S_i|)$ lösen? (1 Punkt)

4. In der Vorlesung wurde zwei Möglichkeiten zur Konstruktion eines generalisierten Suffixbaums für eine Menge von Strings $S_1, S_2, ..., S_m$ diskutiert. Im ersten Fall werden alle Strings, getrennt durch unterschiedliche Trennzeichen, aneinandergehängt. Für diesen langen String wird der Suffixbaum aufgebaut, der dann noch leicht nachbearbeitet werden muss. Alternativ kann man die Strings auch sequentiell mit dem Ukkonen-Algorithmus in einen Suffixbaum integrieren. Dazu wird zuerst für den String S_1, versehen mit einem Endezeichen, der Suffixbaum konstruiert. Zur Integration von S_2, versehen mit dem gleichen Endezeichen, wird nun wiederum der Ukkonen-Algorithmus benutzt. Der Algorithmus wird aber sofort mit Phase $i + 1$ gestartet, wobei i die Länge des längsten Präfixes von S_2 ist, der, von der Wurzel des Suffixbaums ausgehend, im Suffixbaum von S_1 enthalten ist. Für die Strings $S_3, ..., S_m$ erfolgt die Integration analog. Diskutieren Sie auf, warum das Vorgehen korrekt ist. (3 Punkte)