Sommersemester 2006 Abgabe am 16. Juni 2006

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

10. Übung „Algorithmen der Bioinformatik I“

1. Suffixarrays können anstelle von Suffixbäumen zur Suche eines Musters P in einem Text T genutzt werden. In der Vorlesung wurde ein Algorithmus vorgestellt, mit dem bei der Suche nach dem Muster der linke Index des Intervalls, das die potentielle Fundstelle einschließt, bestimmt werden kann.

   a) Stellen Sie einen Algorithmus auf, der auf analoge Weise den rechten Index dieses Intervalls bestimmt. (2 Punkte)

   b) Implementieren Sie die Suche eines Musters P in einem Text T mithilfe eines Suffixarray unter Verwendung dieser beiden Algorithmen zur Bestimmung der Intervallgrenzen. Die Indexe der Intervalle während der binären Suche und die Funstellen (wenn vorhanden) des Musters sollen ausgegeben werden. Der Aufbau des Suffixarrays kann dabei „naiv“ durch eine einfache lexikografische Sortierung der Suffixe (beispielsweise durch vorhandene Methoden zur Sortierung) erfolgen. (5 Punkte)

2. Sei $lcpl_i(T, P)$ (*longest common prefix length*) die Länge des längsten Präfixes von P, das auch Präfix des Suffixes an Position i des zu T gehörigen Suffixarrays ist.

   a) Stellen Sie das Suffixarray für $T = acaaacacat$ auf und bestimmen Sie für jede Position i des Suffixarrays $lcpl_i(T, P)$ für das Muster $P = act$. (2 Punkte)

   b) Gegeben seien zwei Positionen i und j, $i < j$, im Suffixarray für T und $lcpl_i(T, P)$ und $lcpl_j(T, P)$ bezüglich eines Musters P. Stellen Sie eine Hypothese über die $lcpl_k(T, P)$ für alle Indexe $i < k < j$ in Abhängigkeit von $lcpl_i(T, P)$ und $lcpl_j(T, P)$ auf und begründen Sie deren Korrektheit. (3 Punkte)

   c) Wie kann die Hypothese auf Aufgabe b) genutzt werden, um die binäre Suche im Suffixarray zu beschleunigen? (*Hinweis: Was nimmt bei der binären Suche die meiste Zeit in Anspruch?*) (3 Punkte)