

## Aufgabenblatt 8

### Suffixtrees und Suffixarrays

---

Abgabe: **04.01.2005** bis **17 Uhr** über Goya

Maximal: 20 Punkte

Namen und Option nicht vergessen!

---

1. Suche in DNA- und RNA-Sequenzen nach Repeats und Palindromen.

Erklärung: In DNA- als auch in Proteinsequenzen kommen Repeats sehr häufig vor.

Zum einen können kurze oder auch längere Abschnitte von DNA mehrmals in Leserichtung wiederholt werden. Beispiele hierfür sind:

- *tandem arrays*. Kurze Abschnitte werden mehrfach (bis zu tausend-fach) wiederholt, beispielsweise kommt die Sequenz **TTAGGG** am Ende jedes menschlichen Chromosoms bis zu 2000 mal vor. Einige Erbkrankheiten haben ihre Ursache in solchen Repeats.
- *short / long interspersed nuclear sequences* (SINEs / LINEs). Das klassische Beispiel für einen SINE ist die *Alu*-Familie. Alu Repeats sind etwa 300 Nukleotide lang, kommen etwa 300 000 mal im menschlichen Genom vor und machen 5 % des Erbguts aus.

Ein weiterer wichtiger Punkt sind Palindrome in biologischen Sequenzen. Ein *Palindrom* ist eine Sequenz, die rückwärts genauso gelesen werden kann wie vorwärts. Für die Biologie wird der Begriff des *komplementären Palindroms* eingefügt. Nachdem DNA und auch RNA Wasserstoffbrücken zu der jeweils komplementären Base ausbilden, sind auch Repeats innerhalb des komplementären Strangs von Bedeutung. Ein Beispiel für ein komplementäres Palindrom ist die Sequenz **AGCTCGCGAGCT**.

```

5'                3'
AGCTCGCGAGCT
TCGAGCGCTCGA
3'                5'
```

Beispiele für das Vorkommen von komplementären Palindromen sind:

- *tRNA*. Die tRNA faltet sich so, so dass jeweils komplementäre Abschnitte des RNA Moleküls sich aneinanderlagern.
- *Restriktionsenzym-Schnittstellen*. Restriktionsenzyme erkennen Muster auf der DNA. Diese Muster sind meist komplementäre Palindrome. Ein Beispiel ist die Schnittstelle von EcoRI mit **GAATTC**.

**Option 1 (20 P)**

1. Skizziert das Vorgehen, um ein komplementäres Palindrom in einer Sequenz (DNA oder RNA) mit Hilfe von Suffixtrees zu finden. **8 P**
2. Womit beschäftigt sich das Gebiet der Systembiologie? **2 P**
3. Notiert das Suffixarray für den String 'gegenkundgebung\$'. **3 P**
4. Sucht in dem Suffixarray das Pattern 'gegen' mit der besten in der Vorlesung vorgestellten Methode (Binäre Suche mit weiteren Vergleichen ab  $p = \min(p_l, p_r)$ ) und notiert die einzelnen Schritte. **3 P**
5. Aus der Vorlesung wissen wir: „Nicht alles, was mit Suffixbäumen geht, geht mit Suffixarrays.“ Nennt und begründet ein Beispiel! **4 P**

**Option 2 (20 P)**

1. Skizziert das Vorgehen, um Repeats (tandem arrays, SINEs und LINEs, ...) in einer Sequenz (DNA oder RNA) mit Hilfe von Suffixtrees zu finden. **6 P**
2. Implementiert Ukkonen's Algorithmus zur Konstruktion von Suffix-Bäumen und sucht nach Repeats in der gegebenen Sequenz (8-1; auf der Web-seite). Gebt alle längsten Repeats (Mindestlänge 10) aus.  
(*Tip*: die beiden Java-Interfaces `Node` und `TreeNode` könnten für die Implementierung hilfreich sein, siehe <http://java.sun.com/j2se/1.4.2/docs/api/org/w3c/dom/Node.html>)

Der Aufruf soll durch

```
java GruppeX Aufgabe8 -s Sequenz-Datei
```

erfolgen.

Ausgabe:

```
gtcagcgctgac
ttggccaa
gaattc
...
```

**14 P**

**Es muß nur eine der beiden Optionen behandelt werden. Aufgaben aus unterschiedlichen Optionen dürfen nicht gemischt werden!**