KTH, NADA, Henrik Eriksson

2D1320, TENTAMEN I TILLÄMPAD DATALOGI

Lördagen den 17 april 1999 kl 14–19

Maxpoäng tenta+bonus = 50+5. Betygsgränser: 25 poäng ger trea, 35 ger fyra, 45 ger femma.
Resultatet anslås senast 24 april på Nadas anslagstavla.
Hjälpmedel : En algoritmbok.

1. *DNA–automat*

En DNA-molekyl kan beskrivas som ett mycket långt ord i alfabetet CGAT, alltså av typen `GTCTAAGCCCTAAGCG...`

(6p) Konstruera en Knuthautomat som söker efter kombinationen `CCACCGCT` i en DNA-molekyl av en miljon bokstäver. Rita upp den med heldragna framåtpilar och prickade bakåtpilaroch ange den `next`-vektor som definierar automaten!

Finns det i detta fall någon ännu fiffigare sökmetod?

2. *Binärträdsintervall*

(6p) DNA-ord för alla människans gener (något hundratusental) finns i ett binärt sökträd med rotpekaren `root` och man vill att anropet `Intervall("CCA","CGD",root)` ska skriva ut intervallet däremellan, alltså alla gener som börjar med CCA, CCC, CCG, CCT, CGA eller CGC. Ange en effektiv rekursiv tanke för `Intervall`, som direkt skulle kunna programmeras.

3. *Genmanipulation*

DNA-molekylen kopierar oupphörligen sej själv. Ibland sker då en transposition, dvs ett avsnitt faller bort och stoppas in igen på fel ställe. Exempel:

```
abcdefghijklmnopqr -> abcijkldefghmnopqr
```

Transpositionen kan beskrivas med fyra tal, nämligen småstyckenas längder (3,5,4,6 i exemplet). Du ska ge en algoritm för att utföra en sålunda specificerad transposition.

a) Om bokstäverna finns i en abstrakt kö och man har ytterligare en kö till sitt förfogande
(4p) under algoritmens gång.

b) Om bokstäverna finns i en abstrakt stack och man har ytterligare en stack till
(4p) förfogande under algoritmens gång.

4. *DNA-hashning*

I Brookhavendatabanken finns bokstavssekvenser för cirka hundratusen gener, där varje gen består av något tusental bokstäver. För snabb sökning används en hashtabell.
(5p) Föreslå hashfunktion och storlek på hashtabellen.

Vänd!

5. *Kortaste transpositionsvägen*

Människor och möss har stora DNA-avsnitt gemensamma, men dom kommer inte i samma ordning. Detta visar att transpositioner har skett under utvecklingens gång. Man vill gärna veta hur detta har gått till och för det behövs en algoritm som löser följande matematiska problem.

Numrera dom gemensamma avsnitten 12345678 i den ordning som dom kommer hos människan. Hos musen är ordningen annorlunda, till exempel 53428761. Vi söker nu
(8p) den kortaste vägen från den ena ordningen till den andra via transpositioner, till exempel

```
12345678->56123478->53476128->53428761
```

Beskriv en algoritm för detta och ange vilka datastrukturer som behövs!

6. *Gensyntax*

Alla gener börjar med `CTT` och slutar med antingen `TAG` eller `GAG`. Antalet bokstäver i genen är alltid delbart med tre.

(4p) Skriv en grammatik för sådana bokstavsföljder.

7. *Gensortering*

Brookhavendatabasen består av cirka hundratusen gener. Den distribueras som en enda lång fil med genorden i bokstavsordning. På KTH har man lyckats analysera ytterligare tusen gener och lagt till dom sist i brookhavenfilen, helt oordnat. Rektor Flod-
(9p) ström, som är pedant, kräver att den nya filen ska sorteras innan den skickas ut i forskarvärlden. Som datakonsult blir du tillfrågad om bästa sättet att utföra sorteringen. Ge uppskattningar av komplexiteten om man använder
a) quicksort,
b) insättningssortering,
c) den enligt dej bästa metoden.

8. *Abstrakta gener*

Ska en gen representeras som en TEXT eller som en siffervektor med siffrorna 1234 eller kanske som en bitvektor där 00,01,10,11 betyder ACGT? Förklara varför det är
(4p) bättre med en abstrakt datatyp i stället och ange vilka anrop en sådan bör ha.