DATATÄHTI '97

Tietotekniikkakilpailun tehtävät

VASTAAJALLE:

Tehtävät on tarkoitettu vastattavaksi lähteitä apuna käyttäen. Lähteet tulee mainita kunkin vastauksen lopussa. Vastausten ei tarvitse olla erityisen pitkiä ja yksityiskohtaisia. Tärkeää on asian tai käsitteen ymmärtäminen ja sen selkeä kuvaaminen.

Vastaa kuhunkin tehtävään 1–4 erillisille A4-kokoisille vastauspapereille. **Kirjoita oma nimesi ja koulusi nimi jokaisen vastauspaperin oikeaan yläkulmaan.**

Ohjelmointitehtävän vastausohje on tehtävän yhteydessä.

Palauta vastausmateriaali opettajallesi viimeistään 10.2.1997

Huom! MAOL ry pidättää oikeuden käyttää vastausmateriaalia haluamallaan tavalla. Vastausmateriaalia ei palauteta.

1. Mitä merkitystä seuraavilla asioilla on ollut tietotekniikan kehityksessä:

a) Charles Babbagen differenssikone
b) Alan Turingin kehittämä Turingin kone
c) Kurt Gödelin epätäydellisyyslause?

2. Kun on haluttu ottaa yhteys kotimikrosta johonkin toiseen tietokoneeseen, se on tehty yleensä modeemia käyttäen. Modeemien nopeudet eivät kuitenkaan ole kasvaneet samalla tavalla kuin kotitietokoneiden muu suorituskyky. Miksi? Mitä muita tekniikoita voitaisiin käyttää? Vertaile tekniikoiden toteutustapaa, suorituskykyä ja käyttötarkoitusta ja arvioi niiden kustannuksia ja käyttöönottomahdollisuuksia.

3. Markkinoilla on tarjolla useita käyttöjärjestelmiä, joiden ominaisuudet poikkeavat toisistaan. Käyttökelpoisuuden voi varmimmin todeta kokeilemalla. Mitä asioita on käyttöjärjestelmän vaihdossa otettava huomioon ja millaisia ongelmia voidaan tässä yhteydessä kohdata?

4. Internet-tietoverkon käyttökelpoisuuteen vaikuttaa merkittävästi se, miten haluttua tietoa tai palvelua voidaan löytää. Tiedonhakuun on kehitetty useita hakupalveluja, joiden tapaan löytää ja luokitella tietoa hakijan on lähes pakko turvautua. Eri palvelujen hakutulokset saattavat poiketa toisistaan huomattavasti. Miten ja mistä syystä? Kerro hakupalvelujen toimintaperiaatteista ja niistä tekniikoista, joilla haut toteutetaan. Kuvaile hakukielten ominaisuuksia sekä arvioi lisäksi, millaisiin käyttötarkoituksiin erilaiset hakupalvelut soveltuvat ja millaista tietoa palvelujen avulla on vaikeaa tai mahdotonta etsiä.

Mitä tarkoitetaan henkilökohtaisilla hakuagenteilla? Millaisia ne voisivat olla?

5. Ohjelmointitehtävä: Geenipankki

Geenipankissa on talletettuna tietoa eri eliöiden soluista löydetyistä DNA- ja proteiiniketjuista ja erilaisista muista proteiinirakenteista. DNA muodostuu neljästä nukleiinihaposta: adniinista (A), tymidiinistä (T), guaniinista (G) ja sytosiinista (C). Oleellista on se, kuinka paljon ja missä järjestyksessä näitä neljää osasta kulloinkin esiintyy.

Eräs geenitutkimukseen liittyvä osaongelma on löytää tunnetun DNA-ketjun esiintymiä geenipankkiin talletetusta valtavasta tietomäärästä. Ongelmaa vaikeuttaa huomattavasti se, että yleensä ei ole tarpeen löytää täsmälleen samanlaista DNA-ketjua vaan riittää löytää etsittyä DNA-ketjua tarpeeksi muistuttava jakso.

Tehtävänäsi on laatia ohjelma, jolla voidaan mm. etsiä geenipankista annettuja DNA-ketjuja muistuttavia osia. Sekä geenipankki että DNA-ketjut on esitetty merkkijonoina, jotka koostuvat merkeistä "A", "T", "G" ja "C".

Jos geenipankista löytyy täsmälleen etsitty DNA-ketju, niin sanotaan, että saatiin asteen 0 täsmäys. Etsiessä voidaan myös sivuuttaa joitakin geenipankin merkkejä. Jos esimerkiksi geenipankissa on merkkijono TAGCGC ja etsitään merkkijonoa TGGC, niin saadaan täsmäys, jos ohitetaan toisena merkkinä oleva A ja neljäntenä merkkinä oleva C. Kun täsmäys saadaan ohittamalla k merkkiä, sanotaan että löydetty täsmäys on astetta k, esim edellä saatiin astetta 2 oleva täsmäys.

Geenipankki on siis käytännössä merkeistä "A", "T", "G" ja "C" koostuva merkkijono. Geenipankin osapankki on osa geenipankin merkkijonosta.

Seuraavassa on selitetty käytettävien syöte- ja tulostiedostojen sisällön muoto ja samalla se, mitä ohjelmasi pitää tehdä saadakseen syötetiedostoista aikaan halutunlaisen tulostiedoston. Lopussa oleva esimerkki varmaankin auttaa ymmärtämisessä.

Syötetiedostot

Ohjelman syöte koostuu kahdesta tiedostosta. Ohjelman ohjaustiedostossa on ensimmäisellä rivillä haettavien DNA-ketjujen lukumäärä n, $0<n<11$. Seuraavalla n rivillä ohjaustiedostossa on haettavat DNA-ketjut (merkkijonot), kukin omalla rivillään. Haettavan DNA-ketjun maksimipituus on 80 merkkiä. Tämän jälkeen ohjaustiedostossa on operaatioiden lukumäärä m, $0<m<11$.

Seuraavalla m rivillä ohjaustiedostossa on m operaatiota, kukin omalla rivillään. Kukin operaatio on joko "etsi k", "min", "osapankki k" tai "generoi k", missä k on kokonaisluku, $-1 < k < 7$. Operaatioiden merkitykset selviävät tulostiedoston muodon esittelyn yhteydessä.

Ohjelman toinen syötetiedosto on geenipankki, jossa geenipankin sisältö on annettu tekstitiedostossa, jonka rivien maksimipituus on 80 merkkiä. Geenipankin sisältö on käsitettävä yhdeksi merkkijonoksi, vaikka se onkin jaettu riveihin, ts. rivinvaihtojen sijainnilla ei ole merkitystä. Geenipankki voi olla mielivaltaisen kokoinen, eikä sen kokoa ole ilmoitettu tiedoston alussa.

Tulostiedosto

Jokaisen operaation vastaus tulostetaan tulostiedostoon ja tämän jälkeen tulostetaan tyhjä rivi. Tulostiedostoon tulostettavat numerot erotetaan välilyönnillä.

Operaation "etsi k" tuloksena tulostetaan tulostiedostoon rivi, jossa jokaista etsittävää n DNA-ketjua kohti on ensimmäisen löydetyn täsmälleen astetta k olevan täsmäyksen alkukohdan ensimmäisen merkin järjestysnumero geenipankissa. Mikäli vaadittua täsmäystä ei löydy, tulostetaan tämän merkkijonon kohdalle 0.

Operaation "min" tuloksena tulostetaan tulostiedostoon rivi, jossa jokaista etsittävää n DNA-ketjua kohti on tulostettu pienin sellainen k, että merkkijonolle löytyy astetta k oleva täsmäys geenipankista. Jos sellaista lukua k ei löydy, että $k<7$, niin tämän DNA-ketjun kohdalle tulostetaan 0.

Operaation "osapankki k" tuloksena tulostetaan tulostiedostoon rivi, jossa on lyhimmän sellaisen osapankin alkukohta ja loppukohta (tiedoston merkkien järjestysnumeroina), että osapankki sisältää kaikkien haettavien DNA-ketjujen korkeintaan astetta k olevat täsmäykset. Jos lyhimpiä osapankkeja on monta, niin tulostetaan ensimmäisen alku- ja loppukohta. Jos haettua osapankkia ei ole olemassa, tulostetaan alku- ja loppukohdaksi nolla.

Operaation "generoi k" tuloksena tulostetaan mielivaltaisen pitkillä riveillä esitettynä mahdollisimman lyhyt sellainen generoitu geenipankki, joka sisältää kaikkien haettavien DNA-ketjujen korkeintaan astetta k oleva täsmäykset. Tämän tehtävän vastaus ei siis riipu käytettävästä geenipankista vaan ainoastaan haettavista merkkijonoista.

Esimerkki

Ohjaustiedosto:

```
2
AAAA
TTGCA
5
etsi 0
etsi 1
min
osapankki 1
generoi 0
```

Geenipankki:

```
AAAATTACT
TGGCAGGGCC
```

Tulostiedosto:

```
1 0

1 9

0 1

1 14

TTGCAAAA
```

Operaation "etsi 0" tulos on helppo tarkastaa: DNA-ketju AAAA löytyy geenipankista alkaen merkistä 1, mutta DNA-ketjua TTGCA ei löydy geenipankista.

Operaation "etsi 1" tulos on perusteltavissa sillä, että DNA-ketjun AAAA merkistä 1 alkava astetta 1 oleva täsmäys saadaan geenipankin merkkijonosta AAAAT ohittamalla lopussa oleva T ja DNA-ketjun TTGCA merkistä 9 alkava täsmäys saadaan ohittamalla merkkijonosta TTGGCA jompikumpi G.

Edellisten vastausten perusteella operaation "min" vastaus on selvä.

Koska DNA-ketjun AAAA ainoa korkeintaan astetta 1 oleva täsmäys alkaa geenipankin ensimmäisestä merkistä, niin on helppoa tarkastaa myös operaation "osapankki 1" vastaus.

Geenipankki TTGCAAAA sisältää kaikkien etsittävien DNA-ketjujen astetta 0 olevat täsmäykset, ja on selvää, että TTGCAAAA on myös lyhin tällainen geenipankki, eli samalla vastaus operaatioon "generoi 0".

Tehtävän palautus

Toteuta ohjelmasi siten, että kun se käynnistetään, se lukee työskentelyhakemistosta automaattisesti syötetiedostot ja kirjoittaa vastaavat tulostiedostot. Syötetiedostoja on kymmenen paria, ja jokaisessa parissa on ohjaustiedosto ja geenipankki. Ohjaustiedostojen nimet ovat OHJAUS0.TXT, OHJAUS1.TXT,..., OHJAUS9.TXT ja vastaavien geenipankkitiedostojen nimet ovat GPANKKI0.TXT, GPANKKI1.TXT,..., GPANKKI9.TXT. Ohjelman kirjoittamien tulostiedostojen nimet ovat, vastaavasti, TULOS0.TXT, TULOS1.TXT,..., TULOS9.TXT.

Lisää mukaan perustelu siitä, miksi uskot ohjelmasi toimivan oikein. Arvioi ohjelman suorituskykyä (montako kertaa se joutuu lukemaan geenipankin läpi, kuinka esim. tarvittavien vertailuoperaatioiden määrä riippuu geenipankista ja etsittävistä DNA-ketjuista, jne.).

Jos et osaa toteuttaa kaikkia operaatioita, niin ratkaise edes joitakin. Myös osittaisesta ratkaisusta saa pisteitä. Jos et esim. osaa samanaikaisesti vertailla kaikkia DNA-ketjuja käydessäsi geenipankkia läpi, niin parempi on käydä geenipankki läpi kutakin DNA-ketjua kohti erikseen, kuin jättää tehtävä ratkaisematta tältä osin.

Teknisiä vaatimuksia

Ohjelman laitteistovaatimukset esitetään kirjallisessa dokumentissa käyttöohjeen kanssa. Lähdekielinen ohjelma voi olla paperimuodossa tai tekstimuotoisena tiedostona levykkeellä. Mikäli ohjelma tarvitsee jotakin sellaista kirjasto-ohjelmaa, ohjelmatulkkia yms jota ei voi yleisesti olettaa koneessa olevan, tulee se sisällyttää tekijänoikeuksia noudattaen vastauslevykkeelle. Mukana seuraavia tulkkeja ja vastaavia ohjelmia käytetään vain kilpailuohjelman testaukseen ja ne tuhotaan, kun kilpailu on ohi.

Muista kirjoittaa nimesi ja koulusi nimi jokaiseen paperiin ja levykkeen etikettiin.