Kuvatietokanta DNA-molekyylissä

Miki Kallio

23. toukokuuta 2005

Joensuun yliopisto
Tietojenkäsittelytiede
Pro gradu -tutkielma

Tiivistelmä

Molekyylibiologian työmenetelmien edistymisen myötä ollaan pian tilanteessa, jossa on mahdollista käyttää DNA:ta tiedon tallentamiseen. Jo 1990-luvun puolivälistä asti on DNA:ta pyritty käyttämään apuna ratkaistaessa laskennallisia ongelmia. Nykyisin paljon tutkittu sovellusalue on DNA-muisti, jossa käytetään bakteerin DNA:ta tiedon tallentamiseen.

Tässä tutkimuksessa kuvataan eräs tapa tallentaa tietoa DNA:lle käyttäen nykyisiä menetelmiä. Tutkimuksessa käytettäväksi valmistettiin tietokoneohjelma, joka pystyy muuttamaan kuvia DNA-koodeiksi ja DNA-sekvenssejä kuviksi. Tutkimuksessa selvitetään yksityiskohtaisesti DNA:lle toteutettu kuvien koodausmenetelmä. Tutkimuksen molekyylibiologian osuus koostuu kahdesta erilaisesta tavasta käyttää DNA:ta tiedon tallentamiseen. Ensimmäisessä osuudessa *E. coli* -bakteerin plasmidiseen DNA:han tallennettiin koodattuna pieni bittikartta-kuva. Toisessa tutkimuksen osuudessa liukoisessa muodossa olevasta DNA-tietokannasta yritettiin erotella kuvia värien perusteella, käyttäen apuna biotiinilla leimattuja koettimia.

ACM-luokat (ACM Computing Classification System, 1998 version): A.m, F.1, J.3

Avainsanat: DNA-muisti, DNA-laskenta

Esipuhe

Tämä työ on syntynyt soveltamalla kahden eri tieteenalan, tietojenkäsittelytieteen ja molekyylibiologian tietoja. Ilman asiantuntevaa ohjausta en olisi kyennyt suoriutumaan urakasta, joten haluan lausua erityiskiitokset suunnitteluavusta sekä perinpohjaisesta ohjauksesta FT, dos. Sinikka Parkkiselle ja FK Ilkka Poralille Joensuun yliopiston biologian laitokselta.

Tietojenkäsittelytieteen osalta haluan kiittää professori Jussi Parkkista ohjauksesta sekä mielenkiintoisesta lopputyön aiheesta.

Kirjoitustyön tein ollessani Japanissa Toyohashi University of Technologyn vieraana tutustumassa tarkemmin molekyylibiologian ihmeelliseen maailmaan. Haluankin kiittää professori Yo Kikuchia ja hänen työryhmäänsä (Department of Ecological Engineering) tuesta sekä uusista ideoista, joiden avulla työn olisi voinut suorittaa ehkä hieman paremmin tuloksin.

Sisältö

1 Johdanto

Koska Mooren lain loppu näyttää olevan päivä päivältä lähempänä, ovat tutkijat yrittäneet löytää uusia, perinteistä tekniikkaa parempia ratkaisuja tietojenkäsittelyyn. Yksi vaihtoehtoinen menetelmä on käyttää DNA-molekyylejä tiedon käsittelyssä ja tallentamisessa.

Tämä johdantoluku jakautuu kolmeen kohtaan. Kohdassa 1.1 kerrotaan tiedon tallentamisen kehityksestä ja kuvataan tallennustarpeen kasvun syitä, kohdassa 1.2 kerrotaan lyhyesti tiedon tallentamisesta DNA:lle ja kohdassa 1.3 käydään läpi tutkimuksen rakenne.

1.1 Tiedon tallentamisen kehitys

Vuosituhansien ajan ihmisillä on ollut tarve säilyttää tietoa. Jo ennen paperin keksimistä ihmiset tallensivat tietoa kaivertamalla luuhun ja kiveen. Ajan myötä helmitaulua ja paperia korvaamaan haluttiin kuitenkin kehittää tehokkaampia välineitä. Babcockin *& al.*, (2004) mukaan 1800-luvun lopussa keksittiin tyhjiöputki, jota käytettiin varhaisissa tietokoneissa. Vuonna 1947 tapahtui merkittävä edistysaskel kun William Shockley, John Bardeen ja Walter Brattain keksivät transistorin, joka korvasi tyhjiöputken ja lisäsi huomattavasti tietokoneiden luotettavuutta. Yksittäisiin transistoreihin perustuvan tekniikan rajat tulivat kuitenkin pian vastaa, sillä piirien koko kasvoi liian suureksi, jolloin ne tulivat vaikeiksi käsitellä. 1958 Jack Kilby loi ensimmäisen integroidun piirin eli mikropiirin juottamalla kaksi virtapiiriä yhdelle germaniumin palaselle. Yksittäisten transistorien valmistamisen sijasta, tehtiinkin mikropiirit pian valmistamalla useita transistoreita samalle puolijohteen palaselle.

Nykyaikainen mikropiiri esim. mikroprosessori tai elektroninen muisti on Babcockin *& al.*, (2004) mukaan noin neliösenttimetrin kokoinen ja sisältää kymmeniä miljoonia transistoreita. Mikropiirien kehitys on merkinnyt jatkuvaa komponenttien koon pienentymistä, hinnan laskua, luotettavuuden parantumista ja tarvittavan tehon sekä jännitteen madaltumista. Kehitys onkin mahdollistanut jatkuvasti monimutkaisempien elektronisten laitteiden valmistamisen sekä suurempien tietomäärien tallentamisen.

Nykyisen kehityksen mahdollistaneella piioksidilla (SiO_2), on Hakalan (2004) mukaan eristeenä erinomaiset sähköiset ominaisuudet. Tulevaisuuden mikropiirit vaativat kui-

tenkin suurempaa transistorien määrää, jolloin yksittäisten transistorien kokoa on pakko pienentää. Tähän asti kehitys on edennyt varsin suoraviivaisesti, noudattaen Mooren lakia, mutta lähestymme tilannetta, jossa kehitys uhkaa hidastua. Siirryttäessä transistorien valmistuksessa jatkuvasti pienempiin mittasuhteisiin, alkavat piioksidin ominaisuudet käydä riittämättömiksi. Mikropiirien kehityksen jatkamiseksi tarvitaan piioksidin korvaava vaihtoehtoinen materiaali.

Tutkimus uuden materiaalin löytämiseksi on Hakalan (2004) mukaan maailmanlaajuista, mutta tutkimusalue on hyvin laaja ja monin osin huonosti tunnettu. Suuri haasteita on saada tutkimus valmiiksi ja tuote käyttöön teollisuuden kaavailemassa neljän vuoden ajassa. Yksi tämän hetken lupaavimmista materiaaleista on hafniumoksidi (HfO_2). Tutkimus on kuitenkin hyvin haastavaa, sillä huomioon on otettava lukuisia erilaisia tekijöitä, kuten esimerkiksi miten tuhannen celsius-asteen lämpökäsittely muuttaa piin ja hafniumoksidin muodostaman raja-pinnan rakennetta.

Erilaisten piin korvaavien materiaalien löytäminen saattaa kuitenkin osoittautua luultuakin vaikeammaksi. Tulevaisuuden tiedontallennuksen tarpeen täyttäminen voi vaatia kokonaan uudenlaista ajattelua ja myös uudenlaisia tallennusmedioita.

Miksi sitten tarvitsemme uutta tallennuskapasiteettia? Suuri syy on se, että monella alalla tutkimuksen tuloksena syntyneet tietomäärät ovat kasvaneet rajusti. Tulosten saaminen nopeasti voi vaatia tiedon organisoimista tietokantajärjestelmien avulla, sopivien käyttöliittymien ja hakujärjestelmien kehittämistä. Tallennustekniikat ovat kyllä kehittyneet huomattavasti lisäten tallennuskapasiteettia ja tiedonhakunopeutta, mutta Fagerholmin (2004) mukaan nykyinen tallennuskapasiteetti Suomessa riittää vain perustoiminnan hoitamiseen. Esimerkiksi bioalan suurimpia haasteita on pystyä hyödyntämään DNA-siruista saatava suunnaton tietomäärä eli yli teratavun suuruinen tallennusmäärä vuosittain. Tämäkin on vielä pieni määrä verrattuna moniin muihin hankkeisiin. CERN:in tulevan LHC-kiihdytin tuottaa yli 10 petatavua raakamateriaalia vuodessa. Ilmatieteen laitos puolestaan ottaa osaa ilmastojärjestelmämallinnukseen, joka tuottaa 250 teratavua tietoa. Kaikkien näiden ja lukuisten muiden tutkimusten tuottama suunnaton tietomäärä sekä yksityisten ihmisten tiedot pitäisi saada tallennettua turvallisesti pitkiksi ajoiksi ja lisäksi tiedon pitää olla helposti saatavilla tai muuten se on hyödytöntä.

Mutta ei riitä, että tieto on tallennettuna ja nopeasti saatavilla. Liian usein käy niin, että tärkeä tieto tai rakkaat muistot katoavat huonon tallennusvälineen vuoksi. Esimer-

kiksi valokuvien värit haalistuvat tai paperi tuhoutuu. Nykyisin voimme toki tallentaa henkilökohtaiset tietomme vaikka CD-levyille, kuten usein tehdään esimerkiksi digitaalisille kuville, mutta onko käyttämämme formaatti ikuinen? Tuskinpa. Riittää kun muistelemme mitä tapahtui 1970-luvulla markkinoille tulleille 5.25 tuuman levykkeille eli lerpuille. Niiden käyttäminen tänään on lähes mahdotonta. Onko meillä enää kolmenkymmenen vuoden kuluttua välineitä CD-levyjen lukemiseen? Entä sadan vuoden kuluttua? Saatavillamme on kuitenkin edullinen tallennusmedia, jonka perusrakenne ei muutu ja jonka säilyvyys on erinomainen: DNA.

1.2 Tiedon tallentaminen DNA:lle

Bio- ja nanoteknologian menetelmien edistymisen myötä on tarve ja mahdollisuus kirjoittaa tietoa DNA:lle lisääntynyt. Aritan (2004) mukaan keskeisiä sovelluksia tällä hetkellä ovat:

DNA-laskenta, jossa yritetään ratkaista laskennallisia ongelmia molekylaarisen biologian menetelmiä käyttäen.

DNA-leima, jossa käytetään tietyn mittaista oligonukleotidia biomolekyylin tunnistamisessa esimerkiksi cDNA-kirjastosta.

DNA-muisti, jossa käytetään bakteerien DNA:ta pitkäaikaisena, äärioloja kestävänä orgaanisena muistina tiedon tallentamiseen.

DNA-nimikirjoitus, joka on tärkeä esimerkiksi tallennettaessa tekijänoikeustietoja bakteerien genomiin tai haluttaessa piilottaa tekijän tiedot muun informaation sekaan.

Erona perinteiseen bioteknologiaan, edellä kuvatuissa sovelluksissa pyritään tallentamaan synteettisesti tuotettua koodia DNA:han. Koska ongelma pyritään ratkaisemaan DNA:n avulla on tiedon koodausmenetelmä oleellisen tärkeä. Vaikka vuosien varrella on esitetty erilaisia malleja tiedon koodaamisesta DNA-sekvensseihin ja muutamia käytännön sovelluksiakin on toteutettu, ei ASCII-järjestelmän kaltaista standardia ole pystytty luomaan.

DNA-laskennan tutkimus on suuntautunut perinteisesti pääasiassa molekyylien laskentatehoon, mutta luultavasti DNA-tietokoneet eivät pysty koskaan syrjäyttämään perinteisiä menetelmiä tietojenkäsittelyssä. Kuitenkin molekyylibiologian työmenetelmien

kehittyessä, olemme todennäköisesti hyvin pian tilanteessa, jossa tiedon tallentaminen DNA:lle on toteutettavissa kohtalaisen helposti ja pienin kustannuksin.

Vaikka Arita (2004) esittääkin, että DNA-muistia käytettäessä tieto tallennetaan äärioloja kestävään bakteeriin, käytettiin tässä tutkimuksessa *E. coli* -bakteeria, joka on yleinen suolistobakteeri. *E. colin* käyttäminen on siinä suhteessa perusteltua, että sen käytöstä märkälaboratoriotöissä on pitkäaikaisia kokemuksia ja sen kasvattaminen on kohtuullisen helppoa. Luultavasti bakteerit eivät ole paras mahdollinen paikka DNA:n säilymisen kannalta, sillä kuivattu tai pakastettu DNA säilyy hyvin, mutta bakteerien etuna on DNA:n monistuminen bakteerikolonisaation mukana.

Ehkäpä tulevaisuudessa voimme tallentaa henkilökohtaiset tietomme ja vaikka valokuvamme bakteerin DNA:han. Teknologian kehittyessä riittävästi, voimme mahdollisesti ottaa yhden tipan nestettä DNA-tietokannasta, laittaa sen katselulaitteeseen ja ihailla tuon pienen nestemäärän sisältämää suunnatonta tietomäärää.

Tässä tutkimuksessa otetaan ensimmäisiä askeleita tuon tavoitteen saavuttamiseksi, koodaamalla pieniä bittikartta-kuvia DNA:lle ja tallentamalla saadut sekvenssit *E. coli* -bakteeriin sekä nestemäiseen DNA-tietokantaan.

1.3 Tutkimuksen rakenne

Tämä tutkimus koostuu johdantoluvun lisäksi neljästä muusta luvusta. Luvussa 2 kuvataan aikaisempia DNA-laskentaan ja orgaaniseen muistiin liittyviä tutkimuksia. Kohdassa 2.2 kerrotaan lyhyesti miten DNA-laskenta sai alkunsa, sen teoreettisesta taustasta ja käytännön sovelluksista. Kohdassa 2.3 kuvataan DNA:n käyttöä orgaanisena muistina ja kohdassa 2.4 kerrotaan mitä edellytetään onnistuneelta DNA:lle tapahtuvalta koodaukselta.

Luvussa 3 kerrotaan tutkimuksen käytännön toteutuksesta. Kohdassa 3.1 kuvataan lyhyesti tutkimuksen taustaa ja sen eri vaiheet. Kohdassa 3.2 kerrotaan DNA-sekvenssi -ohjelman tarkoitus ja toiminnot. Kohdassa 3.3 kuvataan tutkimuksen molekyylibiologian osuus, se miten koodaus DNA:lle on suoritettu ja miten alukkeet suunniteltiin. Luvun lopussa, kohdassa 3.6 kerrotaan tutkimuksen tulokset.

Luvussa 4 pohditaan DNA-laskennan ja DNA-muistin käyttämiseen liittyviä yleisiä ongelmia. Lisäksi käydään läpi tutkimukseen liittyneitä ongelmallisia tilanteita ja vir-

heiden mahdollisia syitä, sekä esitetään vaihtoehtoisia tapoja ratkaista hankaluuksia aiheuttaneet tutkimuksen osat.

Pro Gradu -tutkielman tekstiosuus päättyy luvussa 5 esitettyyn yhteenvetoon tutkimuksen tärkeimmistä tuloksista.

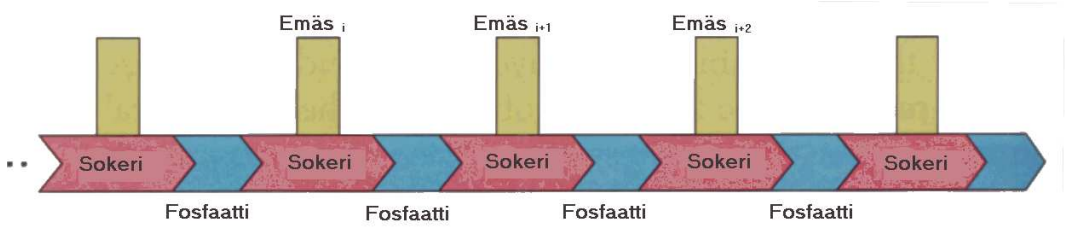
Liitteessä 1 esitellään tutkimuksen keskeiset termit, liitteessä 2 DNA:n sekvensointien tulokset ja liitteessä 3 saatujen sekvenssien sovitukset kuvainformaatiota sisältävien sekvenssien kanssa.

2 DNA:n käyttö tiedon käsittelyssä

Tässä luvussa kuvataan aikaisempia DNA-laskentaan ja orgaaniseen muistiin liittyviä tutkimuksia. Luku koostuu neljästä kohdasta. Kohdassa 2.1 käydään lyhyesti läpi DNA-rakenne. Kohdassa 2.2 kerrotaan miten DNA-laskenta sai alkunsa, sen teoreettisesta taustasta, alalla tehdyistä käytännön sovelluksista sekä siitä millaiset ovat DNA-laskennan tulevaisuudennäkymät. Kohdassa 2.3 kuvataan DNA:n käyttöä orgaanisena muistina ja tietokantahakujen toteuttamismahdollisuuksia sekä kiinteän- että liukoisenfaasin menetelmillä. Kohdassa 2.4 kerrotaan vaatimukset onnistuneelle DNA:lle tapahtuvalle koodaukselle.

2.1 DNA:n rakenne

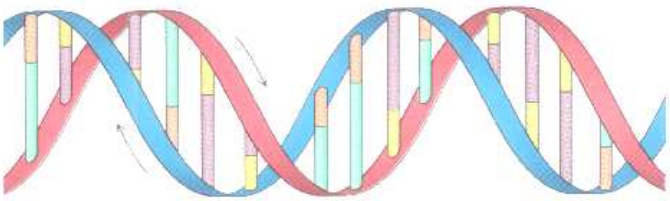
DNA eli deoksiribonukleiinihappo on Bergin *& al.*, (2002) mukaan lineaarinen polymeeri, jonka rakenneyksikkö on nukleotidi. Nukleotidi muodostuu fosforihappotähteestä, sokeriosasta eli deoksiriboosista, josta DNA on saanut nimensä, sekä neljästä erilaisesta orgaanisesta emäksestä, jotka ovat puriiniemäkset adeniini (A) ja guaniini (G) sekä pyrimidiiniemäkset tymiini (T) ja sytosiini (C). Kuvassa 1 havainnollistetaan yksisäikeisen DNA:n rakennetta, jossa vaihtuvat emäkset kiinnittyvät sokeri-fosfaatti -runkoon.



Kuva 1: DNA-säikeen muodostaa sokeri-fosfaatti -runko, johon vaihtuvat emäkset kiinnittyvät (Berg *& al.*, 2002).

Kuvassa 2 on esitetty kuinka DNA:n kaksoiskierteessä (double helix) sokeri-fosfaatti -rungot kulkevat eri suuntiin ulkopuolella ja sisäpuolella on pareittain toisiinsa liittyneet emäkset: {A, C, G, T}. A liittyy aina T:hen ja niiden välillä on 2 vetysidosta. C liittyy puolestaan aina G:hen ja niiden välillä on 3 vetysidosta, jolloin ne sitoutuvat toisiinsa voimakkaammin kuin A ja T. Kyseistä vastakkaisten DNA-säikeiden yhdistymistä kutsutaan Watson-Crick -pariutumiseksi ja sen mukaisesti vastakkaisia säikei-

tä komplementaarisiksi toisilleen. Jokaiselle yksisäikeiselle DNA-sekvenssille kelpaa pariksi vain yhden tietyn emäsrakenteen omaava komplementaarinen säie. Tällöin esimerkiksi ACTTGCAG saa aina komplementaariseksi parikseen TGAACGTC-säikeen.



Kuva 2: DNA:n kaksoiskierre, jossa sokeri-fosfaatti -rungot kulkevat vastakkaisiin suuntiin emästen ulkopuolella (Berg *& al.*, 2002).

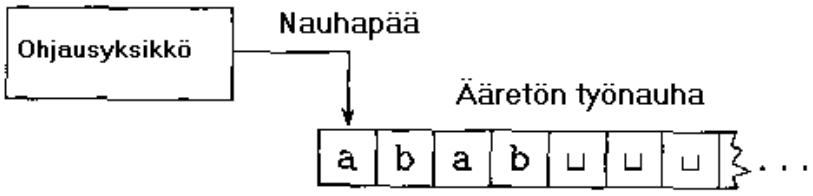
DNA-säikeellä on 5’- ja 3’ -pää t. Emäkset luetellaan aina 5’-päästä lähtien eli esimerkiksi TTATG on eri asia kuin GTATT. 5’-päässä vapaa fosfaatti on liittyneenä deoksiriboosiin. Uudet nukleotidit liittyvät aina DNA-säikeen 3’-päähän eli säiettä syntetisoidaan 5’-päästä lähtien. DNA:n sisältämä geneettinen informaatio sisältyy sen emäsjärjestykseen.

2.2 DNA-laskenta

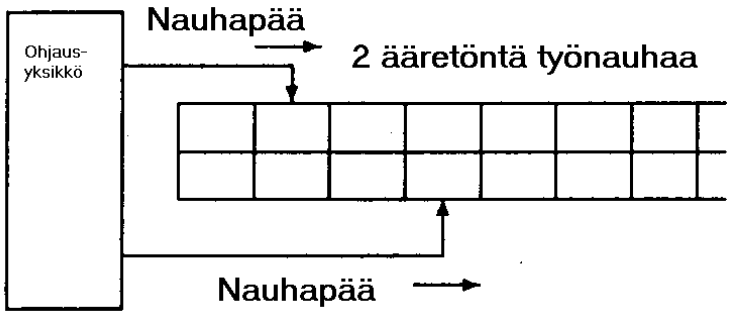
DNA-laskennan voidaan katsoa syntyneen 1994, jolloin Etelä-Kalifornian yliopiston professori Leonard Adleman luki Nobel-palkitun James D. Watsonin kirjaa Molecular Biology of the Gene, ja huomasi DNA-säikeen muistuttavan huomattavasti Turingin konetta. Adlemanin (1998) mukaan hän onnistui ratkaisemaan NP-täydellisen Hamiltonin polku -ongelman käyttämällä laskennassa DNA-molekyylejä. Tuosta uraauurtavasta työstä lähtien ilmaisulla DNA-laskenta on tarkoitettu laskentaa DNA-molekyylien avulla. Ennen Adlemania muutkin olivat tutkineet molekylaarista laskentaa, mutta hän oli ensimmäinen, joka käytti hyväkseen DNA:n mahdollistamaa massiivista rinnakkaisuutta.

Churchin-Turingin -teesin mukaan Turingin koneella voidaan ratkaista jokainen laskennallinen ongelma. Kuvassa 3 on Sipserin (1997) esittämä Turingin kone, jossa on ohjausyksikkö, yksi nauhapää sekä ääretön työnauha, jota voidaan lukea ja kirjoittaa merkki kerrallaan.

Watson-Crick -komplementaarisuudelle perustuva täysin hybridisoitunut DNA-



Kuva 3: Turingin kone, jossa on yksi nauhapää ja ääretön työnauha (Sipser, 1997).

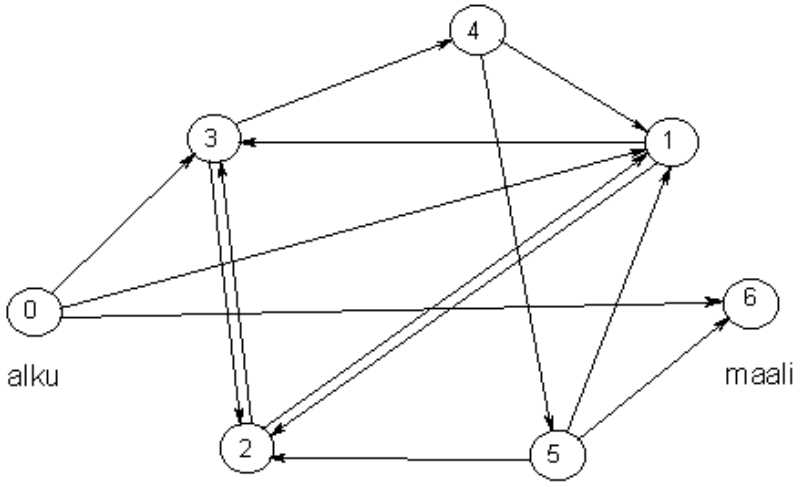


Kuva 4: Watson-Crick -automaatti, jossa on kaksi työnauhaa ja nauhapäätä (Pâun *& al.*, 1998).

kaksoisketju voidaan ajatella merkkijonona, joka koostuu aakkostosta {A, C, G, T}. Onkin varsin luonnollista verrata molekylaarista laskentaa klassiseen formaalien kielten teoriaan. Tietojenkäsittelytieteilijät ovat luoneet malleja erilaisista Watson-Crick -komplementaarisuudelle perustuvista automaateista. Kuvassa 4 on DNA:sta mallinsa saanut Pâunin *& al.*, (1998) esittämä Watson-Crick -automaatti, joka eroaa Turingin koneesta kahden työnauhan ja nauhapään osalta. Periaatteessa kahta nauhaa voidaan käyttää DNA:n kaksisäikeisyyden takia, mutta todellisuudessa tämä on turhaa, sillä säikeet sisältävät komplementaarisuutensa vuoksi saman informaation.

Alkuun DNA-laskennan tutkijat pyrkivät pääasiassa ratkaisemaan NP-täydellisiä ongelmia osoittaakseen uuden menetelmän tehon. Hamiltonin polku -ongelmassa pyritään ratkaisemaan löytyykö annetusta verkosta polkua, joka kulkee täsmälleen kerran jokaisen solmun kautta. Jos verkossa on n solmua, on mahdollisia polkuja n!/2 kappaletta. Tällöin esimerkiksi 13:lla solmulla mahdollisia polkuja on jo 3 miljardia. Adlemanin (1998) mukaan häneltä meni, DNA-laskentaa käyttäen, 7 päivää ratkaista kuvassa 5 esitetty 7 solmua ja 14 kaarta sisältävä Hamiltonin polku -ongelma. Ongelma on tässä mittakaavassa niin helppo, että sen ratkaiseminen paperilla onnistuu keskimäärin

54:ssä sekunnissa, mutta tärkeintä tässä ensimmäisessä työssä oli ratkaisuperiaate.



Kuva 5: Adlemanin Hamiltonin polku -ongelmassa käyttämä verkko (Adleman, 1998).

Adlemanin (1998) mukaan hän käytti seuraavanlaista algoritmia ratkaistessaan DNA-laskennalla n solmuisen Hamiltonin polku -ongelman:

1. Luo joukko satunnaisia polkuja verkon läpi.
2. Käy läpi joukon kaikki polut:
    a) Poista joukosta ne polut, jotka eivät ala aloitussolmusta ja pääty maalisolmuun.
    b) Poista joukosta ne polut, jotka eivät kulje tasan n:n solmun läpi.
    c) Tarkasta jokaisen solmun kohdalta kulkeeko polku sen kautta. Jollei kulje, niin poista polku joukosta.
3. Jos joukko ei ole tyhjä, niin verkosta löytyi Hamiltonin polku.

Algoritmi on toimiva, vaikkakaan ei ole erityisen tehokas. Samanlaista molekylaarisen ohjelmointikielen mallia voidaan soveltaa myös 3-SAT -ongelman (3-CNF) ratkaisuun, josta tuli Hagiyan (2003) mukaan jonkinlainen merkkipaalu yritettäessä testata DNA-laskennan toimivuutta.

Toteutuvuusongelmassa pyritään ratkaisemaan onko annettu lauselogiikan kaava toteutuva, eli tuleeko se todeksi jollain totuusarvovaihtoehdolla. Sipserin (1997) mukaan

3-SAT-ongelma on toteutuvuusongelman erikoistapaus, jossa kaikki kaavat ovat erityisessä muodossa. *Literaali* on muuttuja esim. x tai sen negaatio ¬x, *lause* on joukko literaaleja, joita yhdistää ∧ (disjunktio) tai ∨ (konjunktio). Boolean-kaava on *konjunktiivisessa normaalimuodossa* (CNF), jos se muodostuu ∧:n yhdistämistä lauseista. Kaava on *3-CNF -muodossa*, jos se on konjunktiivisessa normaalimuodossa ja sen lauseissa on korkeintaan 3 literaalia. Esimerkiksi lause $(x_1 \wedge x_3) \wedge (\neg x_4 \vee x_1 \wedge x_2 \wedge \neg x_3)$ ei ole 3-CNF -muodossa, mutta lause $(\neg x_1 \wedge x_2 \vee x_3) \wedge (x_1 \wedge x_2 \wedge \neg x_3)$ on 3-CNF -muodossa.

Hagiyan (2003) mukaan eri tutkimusryhmät käyttivät erilaisia lähestymistapoja: Yoshida ja Suyama ratkaisivat neljän muuttujan tapauksen käyttäen apuna leveyssuuntaista hakua. Sakamoton *& al.*, (2000) mukaan hän ryhmineen käytti hyväkseen yksisäikeisen DNA:n hiusneularakennetta ratkaistessaan 6 muuttujaa sisältävän ongelman ja Hagiyan (2003) mukaan Landweberin johtama ryhmä ratkaisi puolestaan yhdeksän muuttujaa sisältävän tapauksen RNA:n avulla.

Braich *& al.*, (2002) käytti tekniikkaa, joka perustui osasekvenssien erotteluun, jossa käytettiin apuna oligonukleotidikoettimia, jotka oli kiinnitetty polyakryyliamidigeelillä täytettyihin lasimoduuleihin. Tiedon sisältäviä DNA-säikeitä liikutettiin elektroforeesin avulla lasimoduuleiden ohi, jolloin ne säikeet, joissa oli kiinnitetyille koettimille komplementaarisia sekvenssejä, hybridisoituivat kiinni moduuleihin. Kiinni jääneet sekvenssit irrotettiin ajamalla elektroforeesia tarpeeksi lämmitetyissä olosuhteissa, jonka jälkeen vapautetut DNA-sekvenssit ajettiin uusien moduuleiden läpi jne. Tekniikan hyvä puoli on siinä, että laskenta voidaan automatisoida ja sekä lasimoduulit että DNA-sekvenssit voidaan käyttää useampaan kertaan.

Braichin *& al.*, (2002) mukaan laskennan syötteenä käytettiin kahdenkymmenen muuttujan, 24-lauseista, 3-CNF boolean-kaavaa Φ. Jotta tehtävä saatiin mahdollisimman haastavaksi, suunniteltiin se niin, että kaavalla Φ oli vain yksi kelvollinen arvon *tosi* saava ratkaisu. Tehtävässä ei käytetty apuna helpottavia rakenteita, vaan DNA:n avulla käytiin läpi kaikki 2^{20} (1 048 576) mahdollista oikeaa vaihtoehtoa.

Voidakseen suorittaa laskennan Braichin *& al.*, (2002) mukaan työryhmä joutui valmistamaan DNA-kirjaston, jossa oli tarvittavat sekvenssit. He suunnittelivat jokaista kahtakymmentä muuttujaa x_k (k = 1,. . . ,20) varten kaksi selvästi erilaista, 15 emästä pitkää sekvenssiä, joista toinen edusti arvoa *tosi* (T), X_k^T ja toinen arvoa *epätosi* (F), X_k^F. Näiden sekvenssien komplementteja edustivat $\mathcal{X}_k^Z$, jossa k = 1,. . . ,20, Z = T tai F. Jokaista 2^{20} mahdollista vastausta edusti 300:n emäksen kirjastosekvenssi, joka koos-

tui 20:n muuttujan erilaisista yhdistelmistä. Laskennan aikaisten virheiden välttämiseksi käytetyt sekvenssit pyrittiin toteuttamaan siten, että ne eivät olisi hybridisoituneet virheellisesti. Kaikille $\mathcal{X}_k^Z$ sekvensseille syntetisoitiin erottelussa tarvittavat muokatut oligonukleotidikoettimet. Koska pitkien DNA-sekvenssien automatisoitu syntetisointi on hankalaa, jouduttiin pitkät kirjastosekvenssit valmistamaan kahdesta osasta, käyttäen useita eri työvaiheita. Lisäksi kaikkien sekvenssien toiminta oli testattava tarkasti useampaan kertaan ennen varsinaista laskentaa.

Braichin *& al.*, (2002) mukaan tutkimusryhmän käyttämä DNA-tietokone koostui elektroforeesilaatikosta, joka sisälsi kuuman ja kylmän kammion, sekä lasisen kirjastomoduulin, joka puolestaan sisälsi polyakryyliamidigeelissä olevat kovalenttisesti sidotut kaksisäikeiset 300:n emäksen kirjastosekvenssit. Jokaista 24:ää Φ:n lausetta varten koneessa oli lasinen moduuli, joka sisälsi polyakryyliamidigeelissä olevat kovalenttisesti sidotut koettimet, jotka oli suunniteltu kaappaamaan vain ne kirjastosäikeet, jotka koodasivat arvon *tosi* kyseiselle lauseelle. Itse laskenta tapahtui laittamalla kirjastosekvenssit elektroforeesilaatikon kuumaan kammioon ja ensimmäisen lauseen moduuli kylmään kammioon, jonka jälkeen aloitettiin elektroforeesi. Seuraavaksi molemmat moduulit poistettiin laatikosta, kuuman kammion moduuli laitettiin pois käytöstä, laatikko pestiin ja sinne lisättiin uusi puskuriliuos. Kylmässä kammiossa ollut moduuli siirrettiin kuumaan kammioon ja seuraavaa lausetta varten ollut moduuli kylmään kammioon, jonka jälkeen käynnistettiin elektroforeesi. Kuumassa kammiossa olleet kirjastosäikeet denaturoituivat irti koettimista ja siirtyivät kylmään kammioon, jossa lauseen *tosi*-vastauksen omaavat sekvenssit jäivät kiinni moduuliin ja muut sekvenssit jatkoivat matkaa. Samat operaatiot toistettiin kaikille jäljellä oleville lauseille ja lopulta viimeisessä moduulissa oli vain sellaisia säikeitä, jotka olivat saaneet arvon *tosi* kaikille Φ:n lauseille eli itse Φ:n. Tuloksen varmistamiseksi säikeet poistettiin moduulista, monistettiin PCR:n avulla ja sekvensoitiin.

DNA-laskennan avulla on onnistuttu ratkaisemaan useita erilaisia laskennallisia ongelmia, mutta DNA-tietokoneen tehokkuutta on ollut vaikea mitata, vaikka DNA-laskennan kompleksisuus lasketaan kuten muussakin laskennassa, aika- ja tilavaativuuden mukaisesti. Hagiyan (2003) mukaan aikavaativuudella on kaksi puolta: laboratoriossa suoritettujen toimenpiteiden määrä ja kuhunkin toimenpiteeseen käytetty aika. Erityisesti jälkimmäisen huomioiminen on tärkeää analysoitaessa molekylaaristen reaktioiden laskennallista tehoa. Huomioitava on myös se, että mikään kemiallinen reaktio ei tapahdu välittömästi, vaan sen nopeus on aina riippuvainen useammasta eri

tekijästä. DNA-laskennan aikavaativuudesta on julkaistu tutkimuksia ja aikaisemmat tutkimukset keskittyivätkin paljolti erilaisten yksittäisten ongelmien aikavaativuuksien selvittämiseen. Koska tutkimuksissa ei huomioitu todellista ajankäyttöä, ei tuloksilla ole suurta merkitystä. Ilman uusia teknisiä sovelluksia, on DNA-tietokonetta käytettäessä turhaa laskea aikavaativuuksia, sillä manuaalisesti tapahtuvien työvaiheiden osuus on edelleen suuri.

Käytännössä DNA-tietokoneesta on vaikeaa saada nopeaa, sillä koeputkessa tapahtuvat reaktiot ovat usein kovin hitaita. Jos tavoitteena on nopeus, niin DNA-laskennan avulla ei kannata yrittää ratkaista kovin pieniä ongelmia. Toisaalta vaikka DNA-tietokoneilla onkin mahdollista päästä massiiviseen rinnakkaisuuteen, jolla päihittää perinteiset tietokoneet, on Hagiyan (2003) mukaan suurten ongelmien ratkaiseminen todellisuudessa haasteellista. Sopivien, räätälöityjen DNA-sekvenssien, jotka eivät hybridisoidu virheellisesti toistensa kanssa, tuottaminen on vaikeaa, sillä esimerkiksi 50:n muuttujan SAT ongelmaan tarvitaan 1015 erilaista DNA-säiettä. Sekvenssien tuottamista on tutkittu paljon, mutta tällä hetkellä on kuitenkin lähes mahdotonta ratkaista suuria NP-täydellisiä ongelmia käyttämällä DNA-laskentaa.

DNA-laskennassa tilavaativuus on verrannollinen tarvittavien molekyylien määrään, joka ilmentää puolestaan rinnakkaisuuden astetta. Käytettävien molekyylien määrän lisäksi tulee huomioida myös niiden pituus. Yleistämällä voidaan sanoa, että DNA-laskennassa aikavaativuus paranee tilavaativuuden kasvaessa, sillä rinnakkaisuus lisääntyy vastaavasti.

Suurin ongelma DNA-laskennassa on Hagiyan (2003) mukaan virheherkkyys, jota ei voida välttää missään kemiallisissa reaktioissa. Tyypillinen virhe on hybridisaation aikana tapahtuva virheellinen pariutumien, mutta DNA-säie voi myös vahingoittua helposti käsittelyn tai ulkoisten tekijöiden vuoksi. Koska esimerkiksi hybridisaation aikaisten virheiden mahdollisuutta ei voida sulkea täysin pois, on saatujen tulosten analysointi tärkeää kaikissa eri työvaiheista. Virheiden määrän minimoimiseksi on pyritty kehittämään erilaisia, kohdassa 2.4 esiteltäviä, sekvenssien suunnitteluun tarkoitettuja malleja.

2.3 DNA orgaanisena muistina

DNA:ta voidaan käyttää orgaanisena muistina, mutta nykyiset tekniikat rajoittavat sen käytettävyyttä kuten kohdassa 2.3.1 esitetään. Myöskin kohdassa 2.3.2 kuvattavien tietokantahakujen toteuttaminen on nykyisellään vaikeaa.

2.3.1 Orgaaninen muisti

DNA:n käyttäminen orgaanisena muistina on houkuttelevaa, sillä DNA mahdollistaa tiedon tallentamisen erittäin pieneen tilaan. Bergin *& al.*, (2002) mukaan DNA-säikeessä yhden nukleotidin pituus on 3,4Å (0,34 nm) ja leveys noin 10Å (1,0 nm). Näiden tietojen avulla voidaan laskea yhden nanometrin pituisen DNA-sekvenssin tiedontalletuskapasiteetiksi noin 3b ja siitä edelleen 1 cm^2:n, jonka paksuus on 1 nm, tallennuskapasiteetin suuruusluokaksi 10^5 Gb.

Nykyiset tekniikat mahdollistavat vieraiden DNA-molekyylin viemisen vaikkapa bakteerin tai ihmisen soluun. Yleensä tarkoituksena on lisätä tietty DNA-sekvenssi isäntäsoluun jonkin biologisen tutkimuksen puitteissa, ilman tarkoitusta saada jälkikäteen samaa sekvenssiä takaisin solusta. Orgaanista muistia käytettäessä on kuitenkin oleellista saada jo tallennettu tieto myös takaisin muuttumattomana. Harlan *& al.* (2003) esittelee menetelmän, jolla tallennetaan lastenlaulun sanoja plasmideihin, jotka transformoitiin *Deinococcus radiodurans* -bakteeriin. Menetelmä muistuttaa paljolti tässä tutkimuksessa käytettyä tekniikkaa.

Vaikka plasmidiin voidaan liittää vain muutaman tuhannen nukleotidin mittaisia DNA-sekvenssejä, on plamidien käyttäminen perusteltua, sillä nykyisillä tekniikoilla ei voida kuitenkaan syntetisoida kuin noin 100:n nukleotidin mittaisia sekvenssejä. Toisaalta tiedon tallentaminen genomiseen DNA:han ei ole käytännössä toimiva ratkaisu, sillä koodatun tiedon etsiminen genomisesta DNA:sta on lähes mahdoton tehtävä ilman tietoa etsittävästä sekvenssistä. Genomissa tapahtuu myös mutaatioita, jotka vaikeuttavat huomattavasti tiedon säilyvyyttä.

Harlan *& al.* (2003) jakoivat työnsä neljään osaan, joista ensimmäinen oli sopivan isäntäorganismin valinta. Harkittuaan mm. erilaisia kasveja, päätyi työryhmä lopulta valitsemaan kaksi hyvin tunnettua bakteeria: Escherichia coli (*E. coli*) ja Deinococcus radiodurans (*Deinococcus*). Valintaan vaikutti mikro-organismien kyky kasvaa nopeasti

sekä se, että haluttu sekvenssi oli mahdollista eristää niistä nopeasti kokonaisena. Lopullisen tiedon tallennuksen osalta vaaka kallistui *Deinococcucsen* puoleen, sillä se kestää äärimmäisiä olosuhteita kuten UV-valoa, kuivuutta, tyhjiötä, ionisoivaa säteilyä ja joissain tapauksissa myös korkeaa kuumuutta.

Toinen vaihe työssä oli tiedon koodaaminen. Normaalin tietokoneen lukunauhan 0:n ja 1:n sijaan käytössä oli DNA:n neljä emästä A, C, G ja T. Harlan *& al.* (2003) koodasivat englantilaiset aakkoset ja numerot 1 - 9 käyttäen kuhunkin merkkiin kolmen emäksen sekvenssin, jolloin heillä oli käytössään yhteensä 64 erilaista merkkiä.

Kolmas vaihe työssä oli etsiä tunnistussekvenssejä eli isäntäbakteereihin sisältymättömiä sekvenssejä. *E. colin* ja *Deinococcuksen* koko genomi tunnetaan, joten Harlan *& al.* (2003) pyrkivät löytämään 20 emästä pitkiä sekvenssejä, jotka eivät sisältyneet bakteerien genomiin. Ideana oli luoda sekvenssi, joka estää bakteeria mutatoitumasta ja koodaamasta itselleen haitallisia proteiineja. Samalla valitut sekvenssit toimivat tunnistuskohtina koodatun tiedon alussa ja lopussa. *Deinococcuksen* 10 miljardin mahdollisen kandidaatin joukosta löytyi 25 kelvollista sekvenssiä, jotka sisälsivät kolmen emäksen sarjan TAA, TGA, tai TAG. Nämä kolmikot on ns. *stop kodoneita,* jotka kertovat bakteerille, että on aika lopettaa translaatio.

Neljäntenä vaiheena Harlanin *& al.* (2003) työssä oli käytännön toteutus molekyylibiologian laboratoriossa. Ensimmäinen vaihe laboratoriossa oli luoda kaksi toisilleen komplementaarista, 46 nukleotidia pitkää DNA-sekvenssiä. Nämä sekvenssit koostuivat kahdesta aikaisemmin löydetystä bakteerin genomiin kuulumattomasta sekvenssistä, joiden väliin lisättiin 6 nukleotidia pitkä sekvenssi restriktioentsyymin katkaisukohdaksi. Lopuksi luotu sekvenssi kloonattiin plasmidiin.

Seuraavaksi Harlan *& al.* (2003) lisäsivät koodatun tiedon sisältävän DNA-sekvenssin kloonausvektoriin. Vektori puolestaan siirrettiin elektroporaation avulla *E. coli* -bakteeriin. Vektorin annettiin monistua, jonka jälkeen koodatun tiedon sisältämä DNA-säie siirrettiin *Deinococcus*-bakteeriin pysyvää säilytystä varten. Viimeisenä vaiheena oli koodatun tiedon tutkiminen PCR:n ja sekvensoinnin avulla. Tuloksena oli, että kaikki seitsemän syntetisoitua DNA-sekvenssiä, joiden pituudet vaihtelivat 57-99:n nukleotidin välillä, oli onnistuttu tallentamaan seitsemään eri bakteeriin.

Harlan *& al.* (2003) esittää, että ilman uusia teknologisia edistysaskeleitakin, bakteereihin perustuvan muistin kapasiteettia voidaan nostaa dramaattisesti, sillä yhteen mil-

lilitraan nestettä mahtuu 10^9 bakteeria. Potentiaalisena uhkana ovat mutaatiot, joita esiintyy evoluutiossa DNA:n korjausmekanismeista huolimatta.

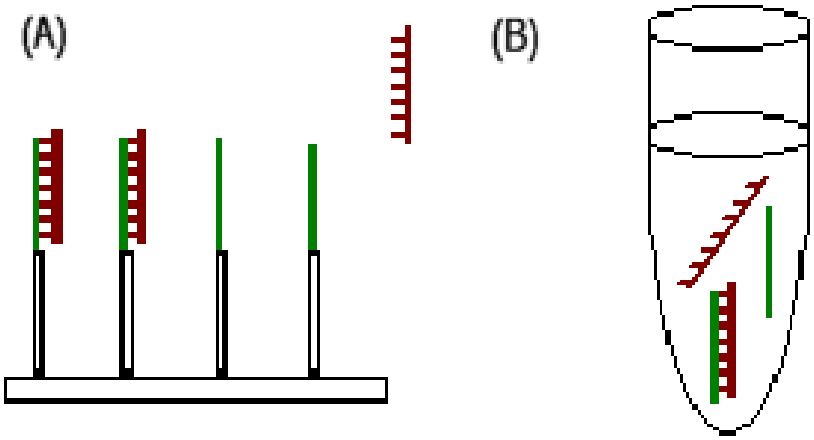
Harlanin *& al.* (2003) työssä koodatussa tiedossa ei esiintynyt mutaatioita, mutta he kasvattivat bakteereita vain sadan sukupolven ajan, mikä on bakteereiden evoluutiossa lyhyt aika. Työssä ei myöskään otettu kantaa mekanismeihin, joiden avulla tietty tieto voitaisiin löytää bakteereiden joukosta. Tämä on kuitenkin mielestäni oleellinen seikka, sillä tallennetun tiedon hyödyntäminen on muuten lähes mahdotonta. Ilman pitkää, jokaiselle tietopaketille yksilöllistä tunnistussekvenssiä, olisi jokainen bakteeripopulaatio kasvatettava erikseen. Ja vaikka bakteereihin voitaisiinkin sisällyttää yksilöllinen tunnistussekvenssi, olisi tiedon etsiminen nykyisillä menetelmillä hyvin hidasta ja vaikeaa.

Ongelmana on lisäksi se, että vaikka käytetty bakteeri kestääkin äärimmäisiä olosuhteita, ei voida olettaa, että se säilyisi missä ja miten pitkään vain. Bakteerille on kuitenkin luotava suotuisat olosuhteet ja sen on saatava ravintoa tai se tuhoutuu. Lisäksi säilytyspaikan on oltava sellainen, että bakteeri on helposti saatavilla. Mielestäni on myöskin harhaanjohtavaa puhua millilitraan nestettä mahtuvasta bakteerimäärästä, sillä vaikka tietoa tallennettaisiinkin miljooniin bakteereihin, olisi niiden joukossa kuitenkin suuri osa täsmälleen samaa informaatiota kantavia bakteereita. 10^9 bakteeria, joissa olisi jokaisessa erilainen tietopaketti on mielestäni käytännössä mahdoton ajatus.

Potentiaalinen paikka tiedon tallentamiseen bakteerien lisäksi on Harlanin *& al.* (2003) mukaan ihmisen omat solut. Omaan DNA:han tallennettu tieto voisikin luoda vaihtoehtoisen tavan tallentaa yksilölle tärkeää tietoa, mutta herättää varmastikin myös keskustelua eettisistä kysymyksistä.

2.3.2 DNA-tietokanta

Mielestäni DNA:lle tapahtuvassa tiedon tallentamisessa ja tietokantahauissa voidaan käyttää periaatteessa kahta erilaista tapaa: kiinteää ja liukoista faasia. Kiinteän faasin -menetelmässä DNA-sanat kiinnitetään kiinteään alustaan kuvan 6 (A) mukaisesti, jolloin saavutetaan kaksi etua: 1) Koska kaksisäikeisen DNA:n toinen säi on kiinnitetty alustaan, voidaan komplementaarinen säi pestä helpommin pois, jolloin pienennetään virheellisen hybridisoitumisen riskiä. 2) Fluoresoivaa leimaa käytettäessä voidaan tunnistaa haluttu sana suuresta tietomäärästä.



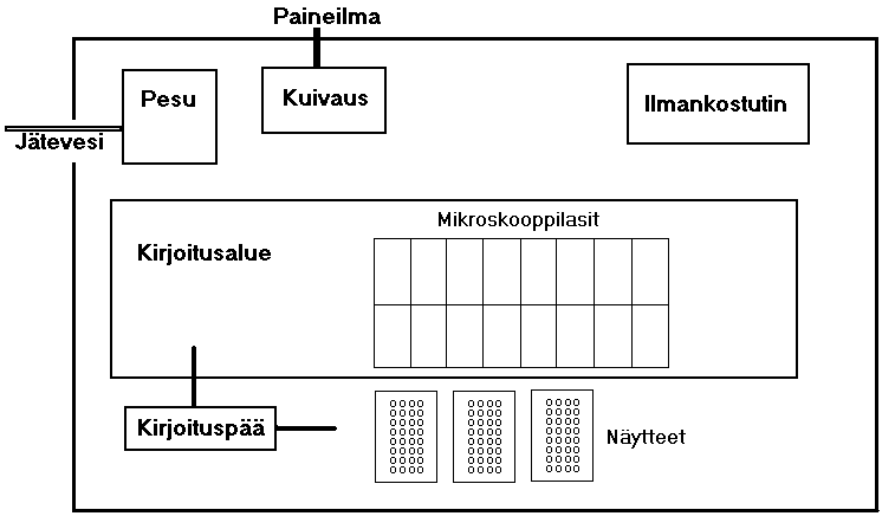
Kuva 6: Kiinteän faasin -menetelmässä (A) DNA-sanat kiinnitetään alustaan, mutta liukoisessa muodossa (B) DNA-sanat liikkuvat vapaasti liuoksessa (Arita, 2004).

Käytettäessä kuvan 6 (B) mukaista, liukoisessa muodossa olevaa DNA:ta, on työn valmistelu helpompaa. Lisäksi etuna on mahdollisuus autonomiseen tiedon prosessointiin, DNA-sanojen lisäämiseen mikrobien genomiin ja nanorakenteiden suunnitteluun.

Tiedon hakumenetelmä pitkäaikaisesta DNA-tietokannasta riippuu ensisijaisesti valitusta tallennusmenetelmästä. Jos tieto on tallennettu bakteerin plasmidiin, on se ensin eristettävä. Jos tieto on ollut tallennettuna kuivattuun tai jäädytettyyn DNA:han riittää pelkkä liukoisen muotoon saattaminen ennen hakua.

Toisaalta jos halutaan tehokasta tietojenkäsittelyä, on jo alkuvaiheessa hyvä tietää millaista tietoa DNA sisältää. Wongin (2003) mukaan DNA-siru antaa kyllä mahdollisuuden käsitellä suuriakin näytemääriä samanaikaisesti, mutta sirujen säilyvyys ei ole kovinkaan hyvä ja niiden valmistus on työlästä. Muita menetelmiä käytettäessä on puolestaan vaikeaa hakea suuria tietomääriä samaan aikaan, joten erityisesti hakujen rajaamisen merkitys korostuu. Toisaalta hakuparametrien suunnittelun vaikeus rajoittaa etenkin suurten tietokantojen tekemistä.

DNA-siru mahdollistaa kiinteän faasin käyttämisen tietokantahauissa. Etuna muihin menetelmiin verrattuna on mahdollisuus käyttää laajaa tietokantaa. Wongin (2003) mukaan siru mahdollistaa kymmenien tuhansien näytteiden rinnakkaisen tutkimisen. Sirujen tuottaminen vaatii korkeaa teknologiaa ja tiedon käsittely voi olla haastavaa, sillä saatavat tietomäärät ovat usein suuria. Sirut mahdollistavat kuitenkin useiden erilaisten biologisten ongelmien ratkaisemisen mm. geenejä tutkittaessa.



Kuva 7: Mikrosirukirjoittimen perusmalli (Wong, 2003).

Normaalisti DNA-sirut koostuvat mikroskooppilasille sijoitetuista tuhansista noin 25-70:n emäksen mittaisista oligonukleotidikoettimista. Wongin (2003) mukaan sirujen tyyppi voi kuitenkin vaihdella tarpeen mukaan. Mahdollisia muita sovelluksia ovat mm. RNA- ja proteiini-sirut. DNA-sirutekniikkaa käytettäessä lasille siirretään yleensä aina kaksi näytettä eli varsinainen tutkittava näyte sekä näyte, johon sitä verrataan. Koska näytteet sijoitetaan lasille tietyssä järjestyksessä on niillä jokaisella oma osoitteensa lasilla ja siten niistä saatu tieto voidaan tulkita jälkikäteen. Sirulla olevien näytteiden määrä voi vaihdella sadoista tuhansiin, mutta yleensä se on useita tuhansia. Siruille ”kirjoitettava” tai niihin hybridisoituva DNA voi olla synteettisesti valmistettuja oligonukleotideja tai PCR:llä valmistettuja DNA-tuotteita. Sirut sisältävät kemiallisesti aktiivisia ryhmiä kuten aldehydejä, jotka auttavat DNA:n stabiloimisessa siruun joko kovalenttisesti tai elektrostaattisilla sidoksilla.

Siruille kirjoittaminen ei ole yksinkertainen tehtävä, vaan vaatii Wongin (2003) mukaan asiantuntemusta kemian, tekniikan, ohjelmoinnin, molekyylibiologian ja projektien hallinnan alalta. Kirjoittamisen aikana on tarkoitus tuottaa kopiointikelpoisia pisteitä johdonmukaisessa järjestyksessä. Nykyisillä kaupallisilla sirukirjoittimilla voidaan valmistaa eri kokoisia ja eri tarkoituksiin soveltuvia siruja ilman erityistä erikoistumista sirutekniikkaan. Kuvassa 7 esiteltävä peruskirjoitin koostuu tärinävaimennetusta pöytäpinnasta, jolle mikroskooppilasit asetetaan, x-y-z -tasoissa liikkuvasta kirjoituspäästä, joka sisältää näytteiden sijoittamiseen tarvittavat neulat ja kynät, niiden pesuun ja kuivaukseen tarkoitetuista asemista, kirjoitusasemasta, sekä tietokoneesta,

joka kontrolloi toimenpiteitä.

DNA-siruja säilytetään tavallisesti valolta suojattuna muovirasioissa. Säilytyslämpötilat vaihtelevat eräiden valmistajien suosittelemasta -20°C:sta toisten suosittelemaan huoneen lämpötilaan. Yleensä sirujen säilyvyyden luvataan olevan noin puoli vuotta.

Fluoresoivan leiman omaavan oligonukleotidin hybridisaatio sirun DNA:han tapahtuu käytännössä samanlaisissa olosuhteissa, kuin normaalistikin molekyylibiologian sovelluksissa. Normaali hybridisaatioliuos sisältää Wongin (2003) mukaan mm. natriumsitraattia (SSC) ja natriumdodekyylisulfaattia (SDS). Hybridisaatiolämpötila vaihtelee käytetystä puskurista riippuen 15-20°C DNA:n sulamislämpötilan alapuolella eli noin 42-50°C:ssa. Tärkeää on pitää lämpötila vakaana ja estää haihtuminen, joten suurilla tilavuuksilla kannattaa käyttää apuna hybridisaatiokammiota. Hybridisaatioon tarvittava aika vaihtelee käytetyn tilavuuden mukaan siten, että pienillä pitoisuuksilla voidaan saada jo muutamassa tunnissa aikaan riittävä hybridisoituminen, vaikkakin yleensä reaktion annetaan tapahtua yön yli. Pesun jälkeen sirut ovat valmiita tulosten lukemista varten. Sirun lukeminen tapahtuu laserin avulla, jolloin sirun DNA kanssa hybridisoituneet, fluoresoivan leiman omaavat oligonukleotidit tulevat näkyviin leimalle ominaisella aallonpituudella.

Liukoisessa muodossa olevan DNA-tietokannan valmistaminen on kohtuullisen helppoa ja nopeaa. Pakastetun tai kuivatun DNA:n säilyvyys on myöskin erittäin hyvä, mutta menetelmän huonona puolena on kerralla haettavan tiedon määrän rajallisuus.

Tiedon hakeminen liukoisessa muodossa olevasta DNA:sta on periaatteessa varsin helppoa. DNA-liuokseen lisätään halutulle säikeelle komplementaarisia yksisäikeisiä oligonukleotideja, joiden 5’-päähän on kovalenttisesti kiinnitetty biotiinileima. Biotiini on molekyyli, joka sitoutuu tiukasti mm. erään bakteerin tuottamaan streptavidiiniproteiiniin. Liuoksen DNA-molekyylit sitoutuvat komplementaarisilta osiltaan vetysidoksin biotiinilla leimattuihin oligonukleotideihin. Liuokseen lisätään paramgneettisia partikkeleita, joiden pinnalle on kiinnitetty streptavidiinimolekyylejä. Biotinyloidut oligonukleotidit kiinnittyvät paramagneettisten partikkelien pinnalle tiukasti, koska niiden päässä oleva biotiini sitoutuu streptavidiiniin. Tuloksena saadaan siis etsityt säikeet sitoutumaan paramagneettisiin partikkeleihin, joissa on rautaoksidia. Partikkeleilla ei ole normaalisti magneettikenttää, kuten raudalla yleensäkään, mutta ne voidaan kerätä talteen magneetin avulla.

Kun liuos viedään magneettikenttään, saadaan paramagneettiset partikkelit ja niihin sitoutuneet DNA-säikeet kerättyä talteen ja muu osa liuoksesta voidaan poistaa. Etsittyjen DNA-säikeiden sekä leimattujen oligonukleotidien väliset vetysidokset pitää vielä katkaista, jotta ne voidaan erotella toisistaan käyttäen jälleen hyväksi magneettikenttää. Lopuksi DNA pitää pestä epäpuhtauksien poistamiseksi, jonka jälkeen tietokantahaku on tehty. Tällä menetelmällä saadaan paljon kiinteän faasin menetelmiä nopeammin eristettyä etsityt DNA-säikeet. Eri toimittajilla on tähän tarkoitukseen valmistettuja reagenssisarjoja ja välineitä.

On olemassa muitakin samantapaisia menetelmiä, joiden avulla oligonukleotidisäikeitä voidaan kiinnittää erilaisiin partikkeleihin, kuten lateksipartikkeleihin. Niiden erottaminen muusta liuoksesta tapahtuu käyttäen erilaisia menetelmiä, mutta perusperiaate on sama.

2.4 Vaatimukset DNA-sanojen suunnitteluun

Hyvä DNA-sanojen suunnittelu on tärkeää, jotta voidaan välttää virheelliset hybridisoitumiset eri sanojen ja niiden komplementtien välillä. Hyvä sanojen suunnittelu mahdollistaa myös tiedon tallentamisen pienempään tilaan sekä suurempien sanajoukkojen muodostamisen.

Koska DNA-sanojen suunnittelu on laskennallisesti vaativa tehtävä, on koodauksessa usein päädytty heuristisiin algoritmeihin, jotka käyttävät esimerkiksi stokastista paikallista hakua.

DNA:lle koodattavat sanat muodostuvat neljästä merkistä: {A, C, G, T}. Aritan (2004) mukaan jokaisen sanan on oltava mahdollisimman erilainen, jotta vältytään virheellisiltä hybridisoitumisilta, olipa sanojen järjestys tai suunta mikä tahansa. Samanaikaisesti sanojen on oltava tasapainossa siten, että biokemiallinen reaktio tapahtuu mahdollisimman tasaisesti.

Vaatimukset sekvensseille: Arita (2004) esittää, että suurimassa osassa koodausmalleista oletetaan DNA-sanojen olevan saman mittaisia. Tämän vuoksi sekvenssien suunnittelussa vastaan tulevat vaatimukset muistuttavat paljolti klassisten virheenkorjauskoodien sanojen suunnittelun vaatimuksia.

Olkoon $x = x_1x_2...x_n$ aakkoston {A, C, G, T} sana. Käänteistä sanaa merkitään $x^R =$

$x_n x_{n-1} ... x_1$ ja komplementaarista sanaa x^C. Hamming-etäisyys $H(x, y)$ kahden sanan $x = x_1 x_2 ... x_n$ ja $y = y_1 y_2 ... y_n$ välillä on indeksien i määrä, joissa $x_i \neq y_i$.

DNA-sanoille voidaan laskea myös käänteinen komplementaarinen Hamming-etäisyys. Aritan (2004) mukaan joukolla DNA-sanoja S, on joukko käänteiskomplementtisanoja S^{RC}, jotka muodostuvat joukon S sanojen käänteisistä, komplementaarista sekvenseistä so. { $x \mid x \in S$ tai $(x^r)^C \in S$}. DNA-sanojen suunnittelu on monimutkaisempaa, kuin tavallisen virheenkorjaavan-koodin. Pelkkä $H(x, y)$:n huomioon ottaminen ei riitä, vaan on myös huomioitava $H(x^C, y^R)$, jotta voidaan välttyä myös komplementtien ja käänteisten sekvenssien aiheuttamilta virheiltä hybridisaation aikana. Perusperiaate on, että käytettyjen DNA-sanojen välillä tulisi olla suuri Hamming-etäisyys.

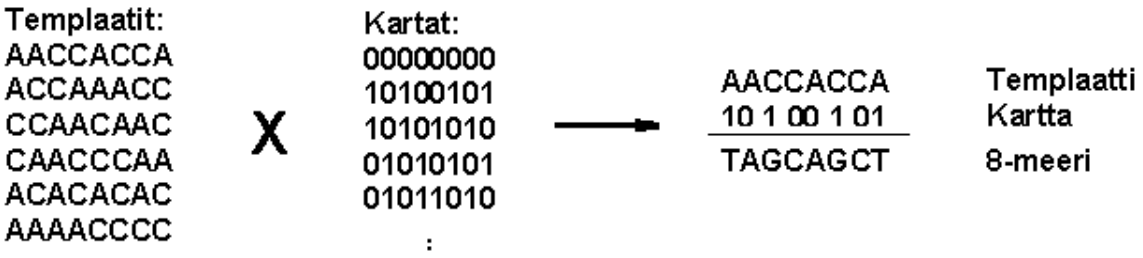
DNA:lla ei ole kiinteää lukukehystä, joten sanojen on oltava vapaita välimerkeistä. Kaikki sanat tulee suunnitella siten, etteivät mitkään kaksi sanan osaa, $x_1 x_2 ... x_n \in S$ ja $y_1 y_2 ... y_n \in S$ (so. $x_{r+1} x_{r+2} ... x_n y_1 y_2 ... y_r; 0 < r < n$) sisälly yhteenkään toiseen joukon *S* sanaan. Jos päällekkäiset sanat eroavat toisistaan ainakin d:n merkin kohdalta, on niiden välimerkitön indeksi d, jonka tulee olla mahdollisimman suuri luku. Tyhjien, ennalta määrättyjen ”aukkojen” lisääminen sanojen väliin ei korvaa lukukehystä, vaikka ne voivatkin helpottaa koodaamista, sillä ne eivät estä virheellistä hybridisoitumista. Ennemminkin tuloksena on koodin pidentyminen, ja sitä kautta käytettävissä olevan informaation määrän pieneneminen.

Toinen tärkeä asia, joka täytyy ottaa huomioon sanoja suunniteltaessa on DNA-sanojen sulamislämpötila (denaturoitumislämpötila), koska sanojen tulee käyttäytyä samalla tavalla *in vitro*. Käytännössä tämä tarkoittaa DNA-sanojen (GC)-pitoisuuden tasaamista. Aritan (2004) mukaan luotettava likiarvo saadaan käyttämällä lähimmän naapurin -approksimaatiota, jossa lämpötila lasketaan emäsdimeerien yleisyyden mukaan. Dimeerit on jaettu kolmeen ryhmään sulamislämpötilan mukaan: [GC][GC] muodostavat vahvimman sidoksen, [GC][AT] tai [AT][GC] seuraavaksi vahvimman ja [AT][AT] heikoimman sidoksen. Dimeerien yleisyys sekvenssissä x on kolmen kokonaisluvun joukko, jossa jokainen luku kuvaa edellä kuvattujen ryhmien yleisyyttä. Sekvenssit tulee ajatella kehämuodossa, jolloin niiden päät yhdistetään. Tällöin esimerkiksi sekvenssien AAGCGCTT ja TACGCGAT sulamislämpötila on lähellä toisiaan, koska niillä on sama dimeerien pitoisuus (3, 2, 3).

Jotta vältyttäisiin virheellisiltä hybridisaatiolta, on kahden edellä kuvatun seikan lisäk-

si suunnittelussa otettava huomioon myös seuraavat seikat: 1) Tekstissä ei saa olla ns. kiellettyjä sanoja, eli osia, joita esiintyy DNA:n katkaisukohdissa, joissa on yksinkertaisia toisto-osuuksia tai muita biologisia signaalisekvenssejä. 2) Missään sanassa ei saa esiintyä k:n (yleensä $k \geq 6$) mittaista sekvenssiä yhtä kertaa useammin. 3) Sanoissa ei saa esiintyä sekundäärirakenteita, jotka estävät halutun hybridisaation eli käytännössä sanat eivät saa hybridisoitua itsensä kanssa muodostaen hiusneularakenteita. 4) Jos käytetään RNA:ta, vain kolmea emästä {A, C ja U}, tulee käyttää suunniteltaessa sanoja.

Templaatti-kartta -strategia on yksinkertainen, mutta tehokas menetelmä DNA-sanojen suunnitteluun. Periaatteena on Aritan (2004) mukaan käyttää kahta binäärikoodia, jotka ovat halutun sanan mittaisia. Templaattia tarvitaan tasoittamaan sanojen GC pitoisuutta ja karttaa generoimaan eroja (mismatch) sanojen välillä. Tuloksena saadaan sana, jossa on mukana kummankin koodin ominaisuudet esim. jos suunnittelu tehdään kuvan 8 mukaisesti, saadaan tuloksena sanoja, joissa on neljä G:tä tai C:tä, ja jotka eroavat toisistaan neljän emäksen osalta.



Kuva 8: Templaatti-kartta -menetelmää käytettäessä sanat saadaan yhdistämällä templaatin ja kartan tiedot, jolloin syntyy 8-meeri (Arita, 2004).

Huonona puolena templaatti-kartta menetelmässä on se, että samoista GC-määristä huolimatta sanojen sulamislämpötilojen erot voivat olla suuret, jopa yli 20 °C. Myöskään sanojen eroavaisuudet eivät riitä aina takaamaan haluttua hybridisaatiota, vaan emäsparien ei-toivotut hybridisaatiot ovat mahdollisia etenkin, jos käytetään suurta sanojen joukkoa.

Stokastisissa algoritmeissa käytetään huomattavissa määrin apuna satunnaisuutta ja niiden avulla onkin pyritty ratkaisemaan mm. tekoälyyn liittyviä ongelmia. Stokastiset paikalliseen hakuun perustuvat algoritmit ovat nykyisin yksi parhaista keinoista ratkaista monia laskenallisesti vaikeita ongelmia.

Stokastinen menetelmä on eniten käytetty tapa DNA-sanojen suunnittelussa. Käytössä

ei ole kuitenkaan standardia, vaan ohjelmia on käytännössä yhtä paljon, kuin on tutkimuksiakin. Deaton *& al.* (2003) käytti geneettisiä algoritmeja löytääkseen sanoja, joilla on sama sulamislämpötila, mutta jotka toteuttavat myös ns. laajennetun Hamming-etäisyyden vaatimukset. Malli eroaa välimerkittömyydestä siten, että siinä huomioidaan suorat sanojen väliset erot emäspareissa, mutta ei sanojen päiden liukumisen aiheuttamia päällekkäisyyksiä. Vaikka stokastisia hakumenetelmiä on käytetty muissakin DNA-sanojen suunnitteluun tehdyissä ohjelmissa, ei algoritmien tehokkuutta ole analysoitu perusteellisesti.

Tulpan *& al.* (2002) esittelemässä stokastisessa, paikalliseen hakuun perustuvassa algoritmissa otetaan huomioon Hamming-etäisyys, käänteinen komplementaarinen Hamming-etäisyys sekä GC-pitoisuus. Hyvänä puolena on lisäksi se, että algoritmin tehokkuutta on pyritty arvioimaan.

Tulpan *& al.* (2002) algoritmi pysähtyy jos kaikki vaatimukset täyttävä sanojen joukko löytyy tai jos asetettu iteraatioiden määrä täytyy. Algoritmin suorituskykyä kontrolloi ns. kohina-parametri, joka määrää algoritmin satunnaisuuden ja ahneuden välisen tasapainon.

Tulpan *& al.* (2002) esittämä DNA-sanojen suunnitteluongelma oli seuraava: syötetään tavoitemäärä *k* ja sanan pituus *n*, etsitään *k*:n suuruinen joukko DNA-sanoja, joiden pituus on *n* ja jotka toteuttavat asetetut vaatimukset. *n* on aakkoston {A, T, C, G} merkeistä koostuva merkkijono, jonka vasen reuna vastaa DNA-säikeen 5'- ja oikea reuna 3'-päätä. Asetettavat vaatimukset ovat jo aikaisemmin tässä luvussa esitellyt: Hamming-etäisyys, GC-pitoisuus sekä käänteinen, komplementaarinen Hamming-etäisyys.

Tietyn mittaisten DNA-sanojen erilaisten variaatioiden lukumäärä on riipppuvainen sanojen pituudesta *n* ja aakkoston merkkien määrästä (4), jolloin variaatioiden määräksi saadaan 4^n. Mahdollisten *k*:n kokoisten erilaisten DNA-sanajoukkojen lukumääräksi saadaan tällöin:

$$\binom{4^n}{k} = \frac{(4^n)!}{k! * (4^n - k)!}$$

Esimerkiksi jos sanan pituus *n* = 8 ja *k* = 100 erilaisten sanajoukkojen lukumääräksi saadaan noin 1.75 * 10^{267}. Luku on niin suuri, että joukkojen etsiminen suoraan laskennallisesti on liian vaativa tehtävä, joten apuna on käytettävä jotain heuristista

algoritmia.

```
procedure StochasticLocalSearch DNA-sanojen suunnitteluun
    input: Sanojen lukumäärä (k), sanojen pituus (n), vaatimusten joukko (C)
    output: Joukko S, jossa m sanaa, jotka täyttävät kokonaan tai osittain vaatimukset C
    for i = 1 to maxTries do
        S = alkuperäinen sanojen joukko
        S' = S
        for j = 1 to maxSteps do
            if S toteuttaa vaatimukset then
                return S
            end if
            Valitaan satunnaisesti sanat w1, w2 ∈ S se. rikotaan yhtä vaatimusta
            M1 = kaikki w1:stä yhtä emästä muuttamalla saadut sanat
            M2 = kaikki w2:sta yhtä emästä muuttamalla saadut sanat
            with probability θ do
                valitse satunnaisesti sana w' joukosta M1 ∪ M2
            otherwise
                valitse sana w' joukosta M1 ∪ M2 se. rikkomusten määrä on minimoitu
            end with probability
            if w' ∈ M1 then
                korvaa w1 w':lla joukossa S
            else
                korvaa w2 w':lla joukossa S
            end if
            if S:llä ei S':a enempää vaatimusten rikkomuksia then
                S' = S
            end if
        end for
    end for
    return S'
end StochasticLocalSearch DNA-sanojen suunnitteluun
```

Kuva 9: Stokastinen paikalliseen hakuun perustuva algoritmi DNA-sanojen suunnitteluun (Tulpan *& al.*, 2002).

Tulpan *& al.* (2002) esittelemässä kuvan 9 algoritmissa optimaalista DNA-sanojen joukkoa etsitään paikallisen haun ja satunnaisten iteraatiivisten parannusten avulla. Kaikissa muodostettavissa sanajoukoissa on täsmälleen sama määrä sanoja. Käytettäessä kriteerinä myös sulamislämpötilaa, voidaan kaikillle muodostettaville sanoille määritellä sama GC-pitoisuus.

Algoritmissa on otettu huomioon myös kaksi muuta laskennallista vaatimusta, Hamming-etäisyys ja käänteinen komplementaarinen Hamming-etäisyys. Tarkoituksena on minimoida mahdolliset ristiriidat kaikkien muodostettavien sanaparien välillä. Alkuperäinen DNA-sanojen joukko määräytyy satunnaisessa prosessissa, jossa generoidaan samat ominaisuudet sisältäviä *n*:n mittaisia sanoja. Sanojen etsiminen voidaan

aloittaa myös käyttäen valmista joukkoa.

Jos joukossa on vähemmän kuin *k* sanaa, voidaan joukkoa täydentää muodostettavilla DNA-sanoilla. Jokaisen iteraation alussa valitaan satunnaisesti sanapari, joka rikkoo yhtä vaatimusta, jonka jälkeen valituista sanoista muodostetaan kaikki mahdolliset yhtä emästä muuttamalla saadut variaatiot.

Parametri θ (kohina-parametri) kontrolloi hakuprosessin ahneutta. Korkeilla θ:n arvoilla vaatimusten rikkomuksia ei ratkota tehokkaasti, kun puolestaan matalilla θ:n arvoilla saavutetaan todennäköisemmin paikallinen maksimi. Paras siihen astisista eli vähiten rikkomuksia sisältävä ratkaisukandidaatti pidetään tallennettuna koko haun ajan. Vaikka algoritmi pysähtyisikin löytämättä kelvollista *k*:n kokoista joukkoa, voidaan kelvollinen osajoukko aina valita iteratiivisen valinnan kautta. Joten kokoa *k* oleva DNA-sanojen joukosta, jossa on *t* kappaletta vaatimusten rikkomusta, voidaan muodostaa ainakin *k* - *t*:n kokoinen kelvollisten sanojen joukko. Sanojen ja niiden komplementtien välistä Hamming-etäisyyttä ei lasketa uudelleen jokaisella iteraatiolla, vaan tilanne päivitetään jokaisen haun jälkeen.

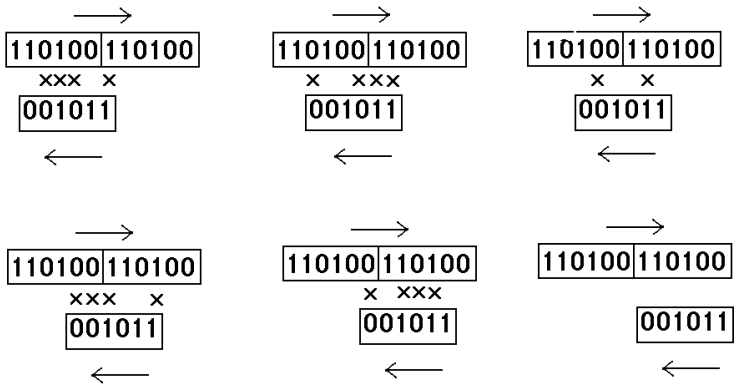
Algoritmin ulompi silmukka voi muodostaa useita itsenäisiä hakuja. Itsenäiset haut voivat yhdessä satunnaisen haun aloittamisen kanssa auttaa saavuttamaan paremman tuloksen, etenkin jos on olemassa vaara haun pysähtymisestä eli jäämisestä alueelle, josta ei todennäköisesti löydy kelvollisia tuloksia. Tällaisissa tilanteissa korostuu parametrin *maxSteps* merkitys, sillä se kontrolloi käytettävien hakujen määrää ja mahdollistaa siten haun aloittamisen uudella sanajoukolla.

Välimerkitön, virheen korjaava DNA-koodi. Vaikeinta DNA-sanojen koodauksessa on Aritan (2004) mukaan välimerkittömyyden saavuttaminen. Ongelmana on, ettei tiedetä systemaattista tapaa luoda korkean indeksin omaavaa välimerkitöntä koodia. Stokastinen haku on puolestaan laskennallisesti liian raskas ollakseen käyttökelpoinen.

Välimerkittömyys on kuitenkin välttämätöntä yleiskäyttöisen DNA-koodin luomisessa. Ratkaisuksi sekvenssien suunnitteluun Arita (2002) esittää ns. *Templaattimenetelmää*, jossa käytetään apuna binääristä templaattia *T* ja virheenkorjaukseen positiivista kokonaislukua *d*. Templaatti *T* valitaan siten, että *T*:n ja seuraavien yhdistelmien välillä esiintyy ainakin *d* kappaletta eroavaisuuksia:

$T^R \quad TT^R \quad T^RT \quad TT \quad T^RT^R$

Missä T^R tarkoittaa käänteistä *T*:tä ja sitä käytetään ilmaisemaan eroja komplementaarisen säikeen ja templaatin välillä. Tuloksena saatu koodi perii ominaisuutensa molemmista binäärisanoista. Tällöin kaikkien sanaparien yhdistelmien tuloksena saadaan sekvenssejä, joissa on vähintään *d* eroavuutta virheenkorjauksen ja templaatin ominaisuuksien mukaan.



Kuva 10: DNA-sanan erot yhdistelmiensä kanssa templaatti-menetelmää käytettäessä (Arita, 2002).

Esimerkkinä voidaan miettiä 6:n nukleotidin mittaisen DNA-sanan suunnittelua. Kuvassa 10 näkyy miten templaatti *T* = 110100 sisältää vähintään 2 eroa *TT*:n ja T^R:n välillä, mutta sama pätee kaikkiin muihinkin edellä kuvattuihin yhdistelmiin.

Templaatti		Koodi		DNA-sana
110100	×	000000	=	CCACAA
		000011		CCACTT
		000101		CCAGAT
		000110		CCAGTA
		001001		CCTCTA
		001010		CCTCTA
		.		.
		.		.
		111100		GGTGAA
		111111		GGTGTT

Kuva 11: Templaatti määrää koodin GC sekä TA -parien sijainnin. DNA-sanojen emäkset saadaan yhdistämällä templaatin ja koodin bitit (Arita, 2002).

Kuvassa 11 on esitetty DNA-sanojen suunnittelu templaatti-menetelmää käyttäen. Aritan (2002) mukaan templaatti määrää DNA-koodin sanojen GC ja TA -parien sijain-

nin. Lopullisten DNA-sanojen emäkset saadaan yhdistämällä templaatin ja koodin bitit. Kuvan osoittamalla menetelmällä saadaan 32 kappaletta kuuden nukleotidin mittaisia sanoja, joilla on vähintään kaksi eroavaisuutta, kun lisätään yksi samanlainen bitti kaikkiin 5:n bitin kokoisiin malleihin.

Saman templaatin käyttäminen on hyödyllistä DNA:n sulamislämmön suhteen. Koska sanojen GC-parien sijainti ja määrä pysyy samana, pysyy myös sulamislämpö samana. Yhtäläinen GC-rakenne auttaa myös käytettyjen sekvenssien etsimistä genomisesta DNA:sta.

Vaikka sulamislämpö on sama ja sanojen välillä on halutun verran eroja (*d*), voivat DNA-sanat silti hybridisoitua virheellisesti keskenään. Siksi on tärkeää, että sanoissa ei ole samoja 7-8 emäksen toistosekvenssejä. Templaatti-menetelmässä toistuvat rakenteet pystytään välttämään erillisellä suunnittelulla, mutta se rajoittaa samalla käytettävissä olevien sanojen määrää. Näistä vaatimuksista huolimatta Arita (2002) pystyi luomaan 112 DNA-sanaa, joiden pituus on 12 nukleotidia, joiden välisten erojen määrä *d*, on vähintään 4. Suunnitelluissa sanoissa tai niiden yhdistelmissä, ei myöskään esiinny yhtään yhtenäistä 7 emäksen mittaista tai sitä pidempää samaa sekvenssiä.

Nykyisin on jo olemassa muutamia ohjelmia DNA-sanojen suunnitteluun. Esimerkiksi DNASequenceGenerator-ohjelman avulla on mahdollista suunnitella DNA-sanoja siten, että sulamislämpötila on huomioitu ja siten, että sanoissa ei ole toistuvia osia.

Tässä luvussa kuvattiin aikaisempia DNA-laskentaan ja orgaaniseen muistiin liittyviä tutkimuksia. Luvussa kerrottin miten DNA-laskenta sai alkunsa, millainen on sen teoreettinen viitekehys ja millaisia käytännön sovelluksia alalla on tehty. Lisäksi luvussa kerrottiin DNA:n käytöstä orgaanisena muistina, DNA-tietokantahauista sekä mitä vaaditaan, jotta voidaan taata onnistunut koodaus DNA:lle. Lopuksi luvussa kuvattiin menetelmiä koodauksen suorittamiseen. Seuraavassa luvussa kuvataan tutkimuksen toteutettaminen käytännössä sekä tutkimuksen eri vaiheiden tulokset.

3 Tutkimuksen toteutus

Tässä luvussa kerrotaan miten tutkimus toteutettiin käytännössä. Luku koostuu kuudesta kohdasta. Kohdassa 3.1 kuvataan tutkimuksen taustaa ja kaikki sen vaiheet lyhyesti. Kohdassa 3.2 kerrotaan DNA-sekvenssi -ohjelman tarkoitus ja toiminnot. Kohdassa 3.3 kuvataan tutkimuksen molekyylibiologian osuus, kohdassa 3.4 miten koodaus DNA:lle on suoritettu ja kohdassa 3.5 miten alukkeet suunniteltiin. Luvun lopuksi, kohdassa 3.6 kerrotaan tutkimuksen tulokset.

3.1 Tutkimuksen tausta ja työvaiheet

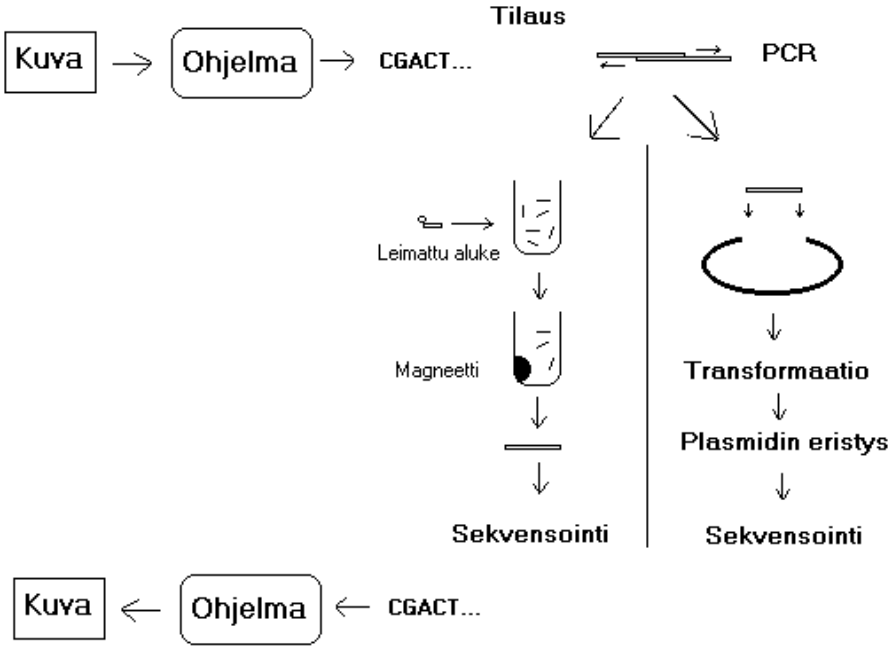
Tutkimuksessa koodattiin DNA-säikeisiin tieto pienien, maksimissaan neljää väriä sisältävien bittikartta-kuvien rakenteesta. Tarkoituksena oli selvittää pystytäänkö *E. coli* -bakteeria käyttämään tiedon tallentamiseen ja pystytäänkö DNA-laskennan menetelmillä löytämään tiettyä väriä sisältäviä kuvia DNA-suspensiosta. Tutkimus jakautui kahteen erilliseen osaan: laboratoriossa tehtyyn molekyylibiologian osuuteen sekä DNA-sekvenssi -tietokoneohjelmaan, jonka avulla pyrittiin todentamaan laboratoriotyön onnistuminen visualisoimalla sekvensoituun DNA:han koodattu tieto.

Kuvassa 12 on esitettynä kaikki tutkimuksen keskeisimmät työvaiheet. Aluksi tutkimusta varten tehtyä DNA-sekvenssi -ohjelmaa käytettiin apuna generoitaessa yksinkertaisista bittikartta-kuvista DNA-sekvenssejä, jolloin tutkittaviin DNA-sekvensseihin saatiin koodattu informaatio kuvien rakenteesta.

Tutkimuksen molekyylibiologian osuus koostui kahdesta erilaisesta tavasta käyttää DNA:ta tiedon tallentamiseen. Molemmissa töissä tilattiin jokaista koodattua kuvaa varten kaksi syntetisoitua aluketta, jotka olivat osittain komplementaariset toisilleen. Kuvia sisältävien DNA-säikeiden valmistaminen ja monistaminen tapahtui näiden alukkeiden ja PCR-laitteen avulla.

Ensimmäisessä molekyylibiologian työvaiheessa kloonattiin T-vektoriin yksi DNA-sekvenssi, jonka jälkeen plasmidi transformoitiin *E. coli* -bakteeriin monistumaan. Kasvatuksen jälkeen plasmidi eristettiin ja sen DNA sekvensoitiin.

Toisessa molekyylibiologian työvaiheessa suoritettiin haku liukoisessa muodossa olevasta DNA-tietokannasta, kalastamalla tiettyjä värejä sisältäviä sekvenssejä DNA-



Kuva 12: Tutkimuksen eri vaiheet: Halutut kuvat koodattiin ensin tietokoneohjelmalla DNA-sekvensseiksi, jotka monistettiin PCR-laitteella ja yritettiin eristää koeputkessa tai transformoida bakteeriin. Eristetty DNA sekvensoitiin ja tulokset tarkastettiin tietokoneohjelmalla.

seoksesta biotiinilla leimattujen alukkeiden avulla kohdassa 2.3.2 kuvatulla tavalla. Lopuksi tietokantahaun tuloksena saatu DNA sekvensoitiin.

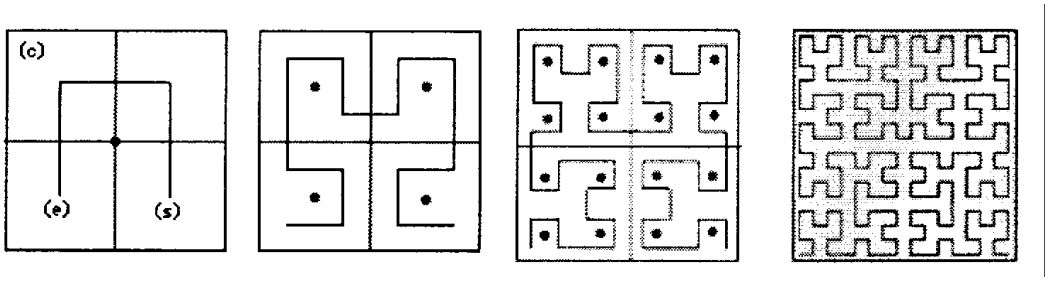
Sekvensoitujen DNA-säikeiden rakenne tallennettiin tekstitiedostona, jonka jälkeen niiden sisältö luettiin DNA-sekvenssi -ohjelmalla. Ohjelman avulla tarkastettiin sisälsivätkö saadut tiedostot koodattua tietoa DNA-sekvenssin sisällöstä. Kelvolliset tiedostot muutettiin DNA-sekvenssi -ohjelmalla 4-värisiksi bittikartta-kuviksi.

3.2 DNA-sekvenssi -ohjelma

DNA-sekvenssi -ohjelmaa käytettiin apuna tutkittaessa laboratoriossa muokattuja DNA-sekvenssejä, joihin oli koodattu informaatiota yksinkertaisista bittikartta-kuvista. Tiettyjä värejä sisältäviä sekvenssejä kalastettiin DNA-seoksesta leimattujen alukkeiden avulla. Yksi sekvenssi insertoitiin plasmidiin ja transformoitiin *E. coli* -bakteeriin, jossa DNA kloonattiin. Plasmidi eristettiin bakteerista ja sen DNA sekvensoitiin, jonka jälkeen sekvensoidun DNA:n rakenne tallennettiin tekstitiedostona.

DNA-sekvenssi -ohjelmalla voidaan lukea bittikartta-kuvia (bmp) ja muuttaa niitä tekstitiedostoiksi, joihin on koodattu tieto kuvan rakenteesta, käyttäen aakkostona DNA:n emäksiä {A, T, G, C }. Vaihtoehtoisesti ohjelmalla luetaan saatuja tekstitiedostoja ja tarkastetaan sisältävätkö ne koodattua tietoa DNA-sekvenssin sisällöstä. Kelvollisia tekstitiedostoja voidaan muuttaa 4-värisiksi bittikartta-kuviksi. Lisäksi ohjelmalla voidaan etsiä sekvensseihin koodattua väriä tai tiettyä kokoa suurempaa värialuetta. Ohjelma soveltuu näiltä osin tarkastamaan laboratoriossa suoritetun työn onnistumista.

Ohjelma hyödyntää kuvien luomisessa ja värialueiden etsinnässä Peano-Hilbert -algoritmia, joka mahdollistaa eri kokoisten, 2^n * 2^n, kuvien lukemisen (n on positiivinen kokonaisluku, joka saadaan vähentämällä kuvan Peano-Hilbertin -asteesta luku 1. 2^n tarkoittaa kuvan x- ja y-akseleiden pituutta pikseleinä). Algoritmi toimii rekursiivisesti aloittamalla lukemisen kuvan oikeasta alanurkasta (piste 2^n, 2^n). Kuvassa 13 on esitettynä Peano-Hilbertin -algoritmin asteet 1 - 4. Kuvan oikeassa reunassa näkyy algoritmin toimintaperiaate 8 * 8 kokoisella kuvalla. Koska tässä tutkimuksessa käytetyt kuvat olivat vain 64:n pikselin kokoisia, muuttaa ohjelma luodut kuvat 400 kertaisiksi suurennoksiksi.

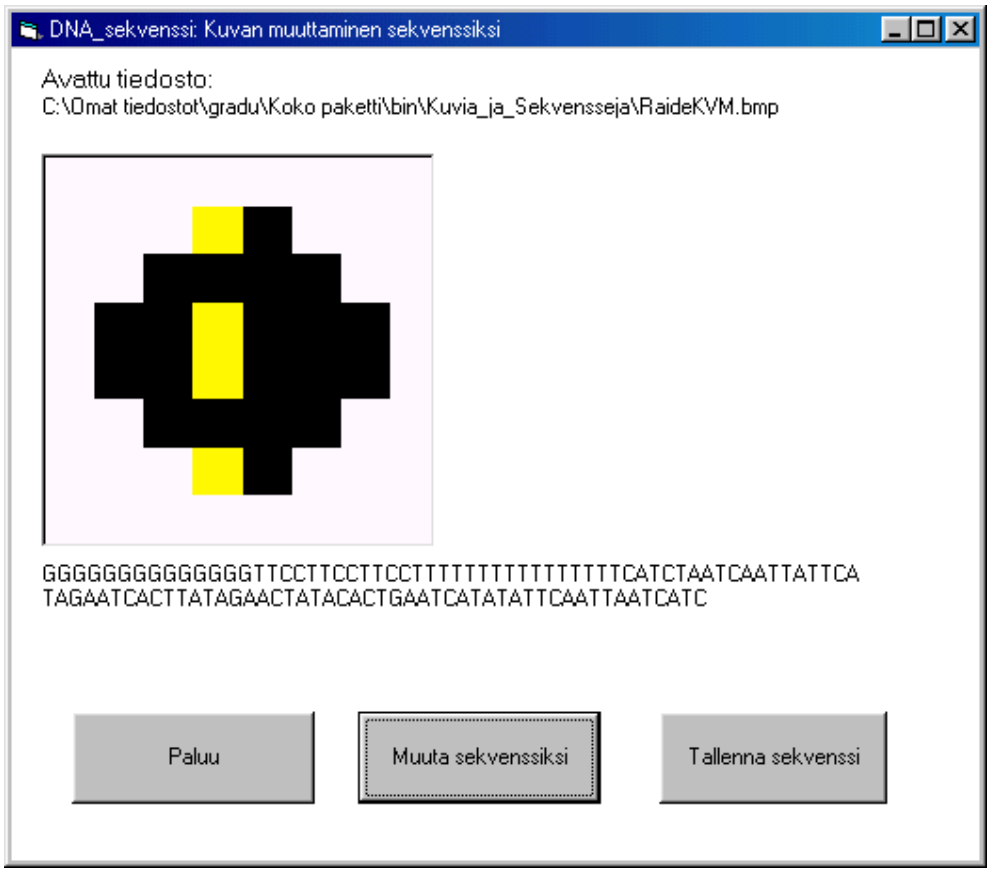


Kuva 13: Peano-Hilbert -algoritmin periaate (Herzog *& al.*, 2002).

Jotta saatuja tietoja ei voitasi hävittää vahingossa, ei DNA-sekvenssi -ohjelman käyttäjälle ole annettu mahdollisuutta muokata tekstitiedostojen sisältöä. Ohjelmassa on viisi päänäyttöä ja yksi erillinen moduuli yleisille vakioille, muuttujille ja aliohjelmille.

Kun DNA-sekvenssi -ohjelma käynnistetään, avautuu Aloitus-näyttö, jolta voidaan valita 4 eri toimintoa: Katsele sekvenssejä, Muuta sekvenssi kuvaksi, Muuta kuva sekvenssiksi ja Etsi värialueita.

Katsele sekvenssejä -näytöllä käyttäjä voi katsella tekstitiedostoiksi tallennettuja DNA-sekvenssejä ja muuttaa niitä niiden koodauksen mukaisiksi 4-värisiksi bittikartta-kuviksi (.bmp). Kuvat käyttäjä voi tallentaa haluamiinsa kansioihin.

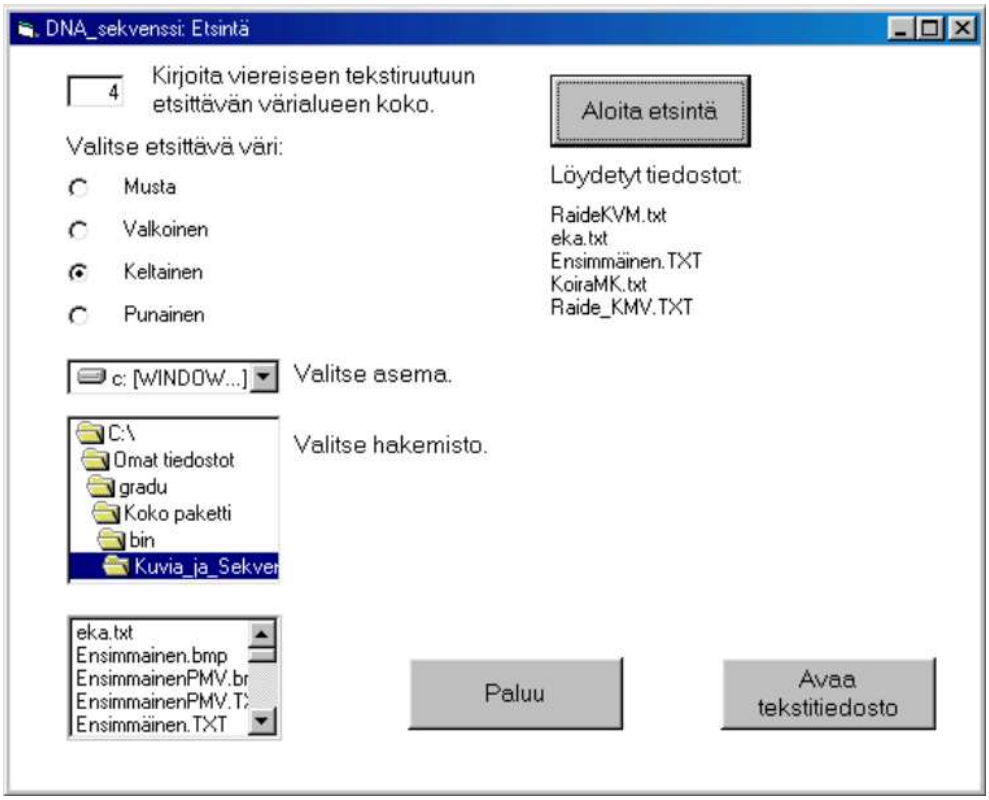


Kuva 14: DNA-sekvenssi -ohjelman avulla tapahtuva bittikartta-kuvan muuttaminen DNA-sekvenssiksi.

Muuta sekvenssi kuvaksi -näytöllä käyttäjä voi muuttaa sekvenssejä niiden koodauksen mukaisiksi 4-värisiksi bittikartta-kuviksi, jotka käyttäjä voi tallentaa haluamiinsa kansioihin.

Muuta kuva sekvenssiksi -näytöllä käyttäjä voi muuttaa kuvan 14 mukaisesti 4-värisiä bittikartta-kuvia DNA-sekvensseiksi. Luodut sekvenssit käyttäjä voi tallentaa haluamiinsa kansioihin.

Kuvan 15 mukaiselta Etsi värialueita -näytöltä käyttäjä voi etsiä koodattuja tiedostoja (DNA-sekvenssejä), jotka sisältävät halutun suuruisia yhtenäisiä värialueita. Ohjelma listaa kansion kaikki määrätyt kriteerit täyttävät sekvenssit, tai antaa ilmoituksen, jos etsintä on tulokseton.



Kuva 15: DNA-sekvenssi -ohjelman avulla tapahtuva yhtenäisten värialueiden etsintä koodatuista DNA-sekvensseistä.

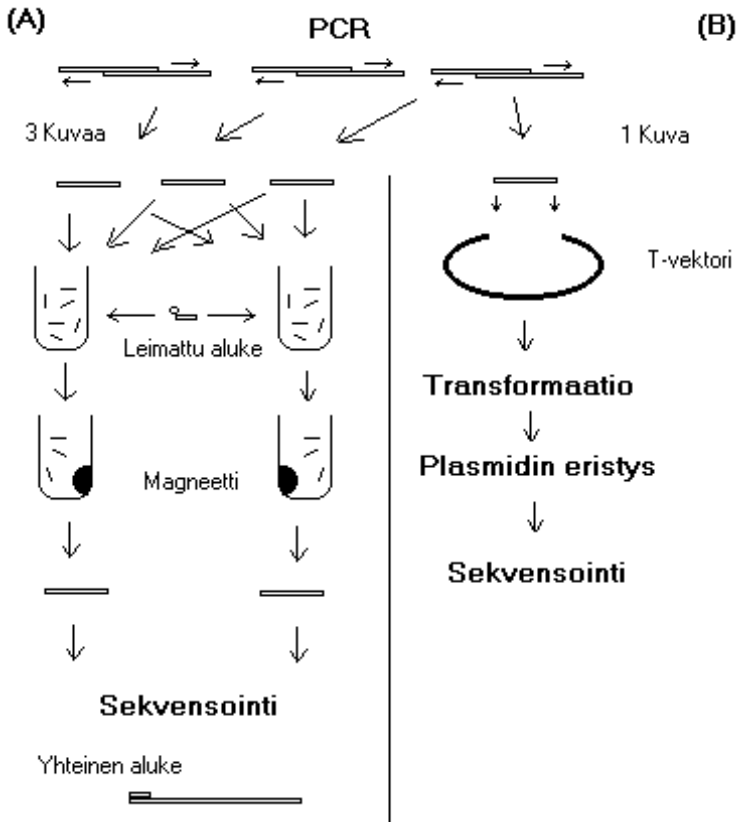
3.3 Molekyylibiologian osuus

Tässä kohdassa kuvataan tutkimuksen molekyylibiologian osuus. Kohdassa 3.3.1 kuvataan tutkimuksen tausta ja tarkoitus, kohdassa 3.3.2 kerrotaan miten tutkimuksessa käytetyt DNA-sekvenssit luotiin ja miten PCR-reaktion onnistuminen tarkastettiin elektroforeettisesti. Kohdassa 3.3.3 kuvataan miten tutkimuksessa käytettiin DNA:ta orgaanisena muistina, kohdassa 3.3.4 kuvataan menetelmät, joilla yritettiin suorittaa tietokantahaku DNA-seoksesta ja kohdassa 3.3.5 kerrotaan sekvensoinnin toteuttamisesta.

3.3.1 Laboratoriotyön tausta

Tutkimuksen molekyylibiologian osuus koostui kahdesta erilaisesta tavasta käyttää DNA:ta tiedon tallentamiseen. Molemmissa osissa valmistettiin PCR:n avulla DNA-säikeitä, joihin koodattiin pienen, maksimissaan 4-värisen bittikarttakuvan rakenne.

Ensimmäisessä tutkimuksen molekyylibiologisessa osassa valmistettu DNA-sekvenssi kloonattiin T-vektoriin, joka transformoitiin *E. coli* -bakteeriin monistumaan. Tämän jälkeen plasmidi eristettiin ja DNA sekvensoitiin, jotta voitiin tarkistaa työn onnistuminen. Kuvan 16 oikeassa laidassa on esitettynä nämä tutkimuksen vaiheet.



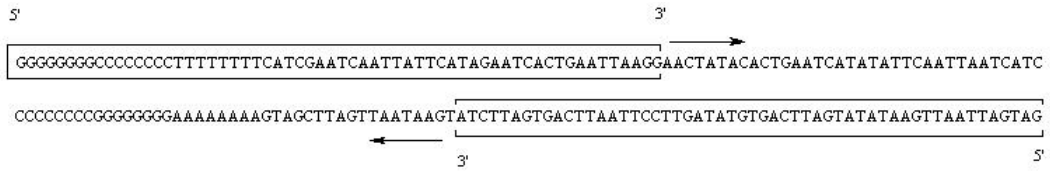
Kuva 16: Tutkimuksen molekyylibiologian osuus jakautui kahteen osaan: (A) PCR:llä tuotettuja sekvenssejä haettiin DNA-tietokannasta biotiinilla leimattujen alukkeiden avulla. (B) Sekvenssi ligatoitiin plasmidiin ja transformoitiin *E. coli* -bakteeriin. Molempien osuuksien tulokset tarkastettiin sekvensoimalla.

Toisessa tutkimuksen osassa luotiin liuokoisessa muodossa oleva DNA-tietokanta sekoittamalla PCR-reaktiosta saadut DNA-säikeet koeputkessa. Tietokannasta suoritettiin haku erottamalla tietyn värin sisältävät säikeet käyttäen DNA-laskennan menetelmiä eli biotiinilla leimattuja koettimia ja magneettia. Tutkimuksessa pyrittiin löytämään seoksesta 2 erillistä kuvaa, joista toinen sisälsi punaista ja toinen keltaista väriä. Molemmissa kuvissa, samoin kuin kolmannessa eli vertailukuvassa, oli myös mustaa ja valkoista väriä. Eristetyt DNA-säikeet sekvensoitiin ja tulokset tarkastettiin tutkimusta varten tehdyllä DNA-sekvenssi -tietokoneohjelmalla. Kuvassa 16 (A) on esitettynä edellä kuvatut tutkimuksen vaiheet.

Molekyylibiologiset työmenetelmät kehittyvät luultavasti muutamassa vuodessa niin, että luvussa 2.4 kuvatut menetelmät voidaan ottaa käyttöön. Niiden mukaisesti koodattujen DNA-sanojen käyttäminen vaatii kuitenkin pitkien sekvenssien hyödyntämistä, johon nykyisin käytössä olevat märkälaboratoriomenetelmät ovat riittämättömiä. Tässä tutkimuksessa jouduttiin tyytymään pitkälti kompromisseihin, jotta käytössä olevien laboratoriomenetelmien avulla voitiin toteuttaa tehokas koodaus DNA:lle.

3.3.2 Polymeraasiketjureaktio

Tutkimuksessa käytetyt DNA-sekvenssit luotiin käyttäen polymeraasiketjureaktiota (PCR), jonka jälkeen reaktion onnistuminen tarkastettiin elektroforeettisesti. PCR:ssä hyödynnetään Bergin *& al.*, (2002) mukaan yksisäikeisen DNA:n kykyä hybridisoitua vastinsäikeensä kanssa. PCR:n aikana toistuvat säännöllisesti kuumennus- ja jäähdytysvaiheet. Kuumennusvaiheen aikana vastinsäikeet irtoavat toisistaan. Kun seosta jäähdytetään alkavat DNA-säikeet jälleen hybridisoitua vastinsäikeidensä kanssa. Tässä vaiheessa on tärkeää, että liuokseen on lisätty tarpeeksi kahta aluketta, jotta DNA:n syntetisointi mahdollistuu. Entsyyminä toimivat DNA-polymeraasit, jotka liittävät syntetisoitavaan ketjuun, 5’-päästä alkaen, templaatin mallin mukaan, yhden nukleotidin kerrallaan. Uusi nukleotidi liitetään aina syntetisoitavan säikeen 3’-päähän. Muodostuva uusi säie on komplementaarinen templaatille. ”Kopiokoneena” toimii PCR-laite ja kopioiden muodostuminen tapahtuu eksponentiaalisella nopeudella, joten tekniikalla saadaan muutamassa tunnissa miljoonia kopioita alkuperäisestä säikeestä. Huomioitavaa on kuitenkin, että kyseessä on monistaminen, eikä tekniikkaa voida käyttää ilman tietoa DNA:n emäsjärjestyksestä.



Kuva 17: Tutkimuksessa käytetyt DNA-sekvenssit luotiin käyttämällä PCR-laitetta. Kaksi toisilleen osittain komplementaarista aluketta mahdollistivat sekvenssien luomisen ilman erillistä templaattia.

Tutkimuksessa käytettiin PCR:ää apuna sekä tutkimuksen alussa haluttujen sekvenssien luomisessa että lopussa monistettaessa sekvensoitavaa DNA:ta. Tutkimuksen alussa tarvittavat DNA-sekvenssit luotiin käyttämällä kuvan 17 mukaisesti kahta toisilleen

osittain komplementaarista aluketta. Kun reaktio käynnistettiin, hybridisoituivat alukkeet keskenään ja Taq-polymeraasin vaikutuksesta kumpikin säi piteni käyttäen toista mallinaan.

PCR:n jälkeen sekvenssit tutkittiin geelielektroforeesilla, jotta voitiin olla varmoja työn onnistumisesta. DNA-molekyylien koko ja puhtaus voidaan määrittää elektroforeettisesti. Nukleotidit ovat negatiivisesti varattuja ja lineaarisen DNA-molekyylin liikkuvuus sähkökentän positiivista napaa kohden on suhteessa sekvenssin pituuteen (molekyylipainoon). Erottelussa käytettävän agaroosigeelin tiheys valitaan molekyylin koon mukaan, sillä lyhyet molekyylit erottuvat paremmin tiheässä geelissä. DNA voidaan havaita geeliltä etidiumbromidilla, joka sitoutuu DNA:han ja fluoresoi ultraviolettivaloa.

Agaroosigeeli sisältää epäpuhtauksia, jotka täytyy poistaa DNA:sta, jotta myöhemmät reaktiot eivät inhiboituisi. DNA-molekyylien puhdistamiseen agaroosigeeliltä on olemassa useita hyviä puhdistussarjoja, jotka sisältävät kaikki tarvittavat materiaalit ja tarkat ohjeet. Valintaan vaikuttaa pääasiassa halutun saaliin määrä ja puhtausaste sekä laboratoriossa tarjolla olevat menetelmät. Tässä tutkimuksessa käytössä ollut QIAEX II Agarose Gel Extraction Protocol on suunniteltu 40 emäsparista (bp) 50 kiloemäksen (kb) pituisten DNA-fragmenttien puhdistamiseen.

3.3.3 *E. coli* orgaanisena muistina

Tutkimuksessa käytettiin *E. coli* -bakteeria orgaanisena muistina, johon tallennettiin luodut DNA-sekvenssit. PCR:n avulla tuotetun DNA-säikeen kasvattaminen ja monistaminen *E. coli* -bakteerissa vaati useita työvaiheita. DNA oli ensin ligatoitava plasmidiin ja plasmidi puolestaan transformoitava bakteerin sisälle. Jotta voitiin olla varmoja, että bakteeri sisältää halutun sekvenssin, oli kolonisaatiosta tutkittava plasmidissa olevan insertin koko geelielektroforeesin avulla. Jotta voitiin eristää riittävä määrä plasmidia sekvensoitavaksi, kasvatettiin bakteereita vielä liuoskasvatuksella.

Tutkimuksessa käytettiin Promegan kaupallista pGEM-T-easy -vektoria, joka hyödyntää PCR:ssä käytetyn Taq-polymeraasin taipumusta liittää säikeen 3’-päähän ylimääräinen deoksiadenosiini. Valmiiksi linearisoidun vektorin 3’-päihin on liitetty yksittäiset tymiinit, joten insertin ligaatio tehostuu huomattavasti, eivätkä vektorin päät siksi kykene juurikaan liittymään toisiinsa ilman inserttiä.

Vektorissa kloonausalue on sijoitettu β-galaktosidaasigeenin α-alayksikön keskelle. Tällöin insertin sisältävien bakteerikloonien pitäisi erottua valkoisina pesäkkeinä, kun taas itseligatoituneen vektorin sisältävät kloonit värjäytyvät yleensä β-galaktosidaasiaktiivisuuden ja maljassa olevan värisubstraatin (X-gal) ansiosta sinisiksi.

Tarkasteltaessa yön yli kasvatettuja bakteerimaljoja, voitiin todeta bakteerikasvuston värin perusteella, että kaikki pesäkkeet eivät sisältäneet luultavasti inserttiä. Liuoskasvatusta varten valittiin kuitenkin 4 pesäkettä, joiden oletettiin sisältävän, sekä 2 pesäkettä, joiden ei oletettu sisältävän inserttiä.

Valittujen pesäkkeiden bakteereita oli tarkoitus kasvattaa plasmidieristystä varten, mutta samaan aikaan haluttiin myös tarkastaa agaroosigeelielektroforeesin avulla, mitkä pesäkkeet sisälsivät insertin. Geelielektroforeesia varten insertit piti ensin monistaa PCR-laitteella käyttäen plasmidille suunniteltuja alukkeita. Alukkeet mahdollistavat DNA:n monistamisen insertin alueelta siten, että vain pieni osa plasmidia tulee mukaan.

Geelielektroforeesin jälkeen voitiin todeta, että kaikkien näytteiden bandit olivat edenneet yhtä pitkälle. Tämä oli yllättävää, sillä bandien joiden ei oletettu sisältää inserttiä, olisi pitänyt kevyempinä liikkua pidemmälle. Koska ei voitu olla varmoja onko insertin ligaatio plasmidiin onnistunut vai sisälsivätkö kaikki pesäkkeet insertin, päätettiin eristää ja sekvensoida yksi molemmista erilaisista pesäkkeestä saatu plasmidi. Plasmidieristys tehtiin Machery-Nagel NucleoSpin Plasmid -menetelmää käyttäen.

Työn lopussa saadut DNA-sekvenssit sekvensoitiin, mutta koska tuloksena oli saatu vain erittäin pieni määrä DNA:ta, täytyi se monistaa käyttäen kohdassa 3.3.5 kuvattavaa sekvensointi-PCR -reaktiota.

3.3.4 Tietokantahaku DNA-seoksesta

Seuraavaksi tutkimuksessa oli vuorossa kuvan 16 (A) mukainen DNA:lle koodattujen kuvien hakeminen liukoisesta DNA-tietokannasta.

Tiettyä väriä sisältävien kuvien etsintä päätettiin toteuttaa käyttämällä oligonukleotideja, joihin oli koodattu haettavalle värille komplementaarinen sekvenssi. Oligonukleotidit olivat biotiinilla leimattuja, jolloin voitiin käyttää apuna paramagneettisia, biotiini-

leimaan kiinnittyviä Dynabeadeja.

Tutkimuksessa käytetyt Dynabeads M-280 Streptavidinit ovat pieniä polystyreenipalloja, joihin on liitetty kovalenttisesti streptavidinia. Ne on suunniteltu sitomaan helposti pieniä biotinyloituja molekyylejä esim. DNA:ta. Dynabeadien kyky sitoa molekyylejä on riippuvainen molekyylien koosta.

Työssä sekoitettiin kahteen mikrosentrifugiputkeen (PUN ja KEL) kutakin PCR:llä aikaisemmin valmistettua DNA-säiettä (DNA:lle tallennetut kuvat). Putkeen KEL lisättiin keltaisen värin komplementtia edustavaa, biotiinilla leimattua aluketta ja putkeen PUN punaisen värin komplementtia edustavaa, biotiinilla leimattua aluketta. Mukaan lisättiin aikaisemmin pestyjä Dynabeadeja ja annettiin säikeiden hybridisoitua.

Seuraavaksi putket siirrettiin magneettialustalle ja niiden annettiin olla paikallaan, jolloin biotiinilla leimatut alukkeet, niihin hybridisoitunut DNA ja biotiinileimaan kiinnittynyt paramagneettiset Dynabeadit tarttuivat magneetin vaikutuksesta putken reunaan. Ennen sekvensointia pelletti pestiin, jotta saatiin poistettua ylimääräinen DNA, jonka jälkeen molemmat näytteet sekvensointiin kohdan 3.3.5 mukaisesti, käyttäen sekvensseille yhteistä sekvensointialuketta.

3.3.5 DNA:n sekvensointi

Sangerin dideoksi-menetelmä on entsymaattinen menetelmä DNA:n sekvensointiin. Se on nykyisin ylivoimaisesti käytetyin tapa sekvensoida DNA:ta.

DNA:n sekvensointi eli emäsjärjestyksen määrittäminen on tärkeä osa geenitekniikkaa, sillä esim. löydetyn, uuden geenin toiminnan tutkimiseksi täytyy ensin määrittää sen emäsjärjestys. Sekvensoinnin avulla voidaan tarkastaa myös valmistettujen DNA-säikeiden virheettömyys. Tässä tutkimuksessa sekvensointia käytettiin varmistamaan tulosten oikeellisuus. Sekvensointi pitää aina suunnitella huolellisesti, sillä huonosti toteutetulla sekvensoinnilla voidaan pilata muuten hyvin onnistunut työ.

DNA:n sekvensointiin on käytettävissä Brownin (2001) mukaan periaatteessa kaksi lähes samanaikaisesti kehitettyä menetelmää: kemiallinen (Maxam-Gilbert) ja yleisemmin käytetty entsymaattinen (Sanger). Menetelmät ovat hyvin erilaiset, mutta ne molemmat mahdollistavat useiden kiloemästen mittaisten DNA-säikeiden nopean sekvensoinnin. Kun geeni on kloonattu plasmidiin ja plasmidi on puhdistettu, on sekvensointi

nykyisin varsin helposti toteutettavissa käyttäen automatisoituja järjestelmiä.

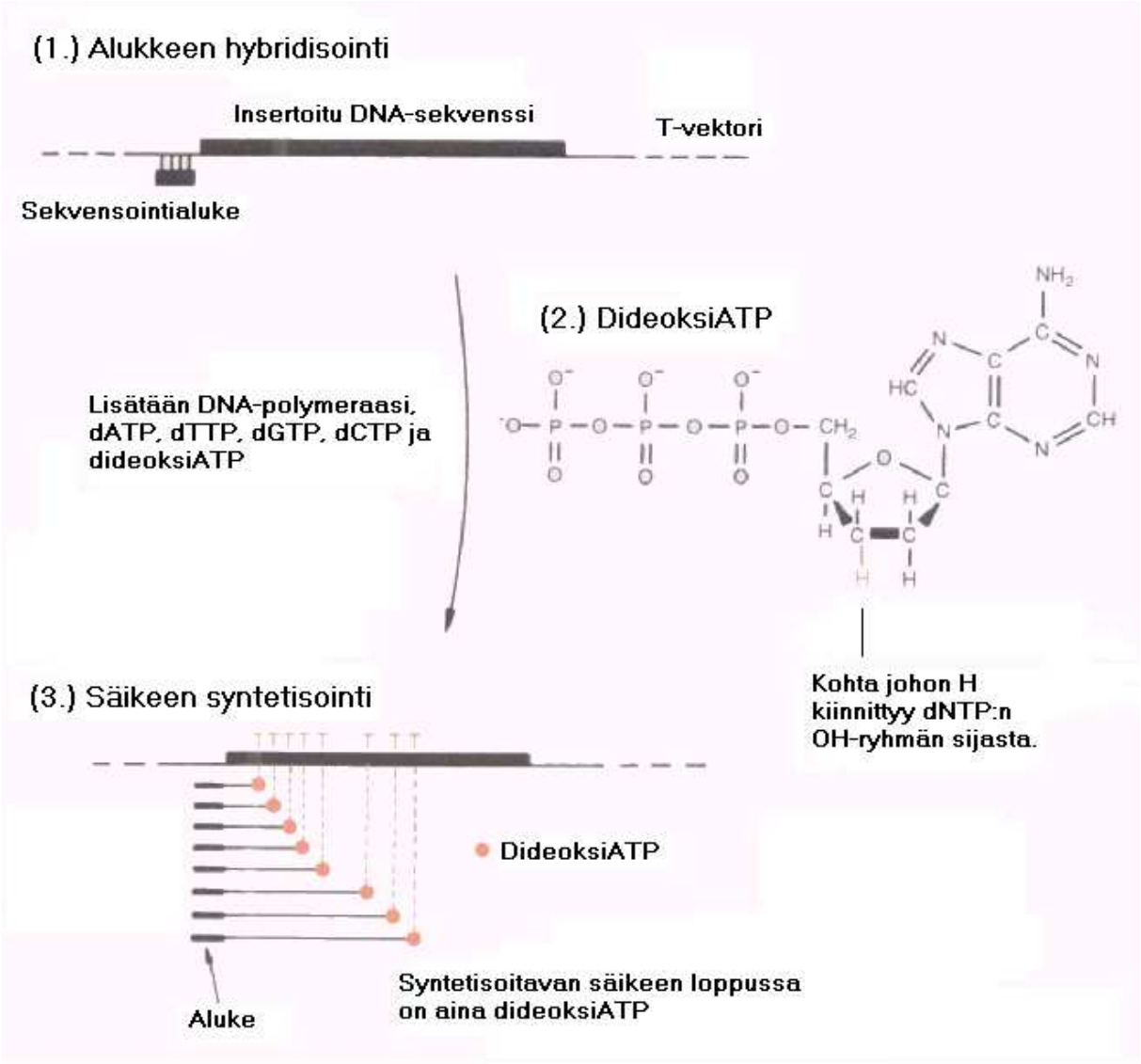
Brownin (2001) mukaan Sangerin menetelmällä voidaan sekvensoida yksisäikeistä DNA:ta. Sekvensoitava molekyyli kloonataan yleensä vektoriin. Lisäksi menetelmässä tarvitaan entsymaattisia synteesejä, jotta voidaan valmistaa templaatille komplementaarinen DNA-säi. Ensimmäinen tehtävä Sangerin menetelmässä on hybridisoida oligonukleotidialuke templaattiin (Kuva 18(1.)). Alukkeen tehtävänä on toimia aloituskohtana komplementaarisen säikeen synteesissä, jonka tapahtuu esim. DNA-polymeraasi I:n Klenow-fragmentin tai muun entsyymin vaikutuksesta. Säikeen syntetisoinnin aloittamiseksi tarvitaan rakennusaineiksi myös kaikkia neljää deoksinukleotidia. Lisäksi mukaan pitää lisätä dideoksinukletidejä (ddNTP), jotka voivat kiinnittyä syntyvään polynukleotidisäikeeseen kuten muutkin nukleotidit. Dideoksinukleotidin kiinnittyminen lopettaa kuitenkin säikeen pitenemisen, sillä dideoksinukleotidilla ei ole vapaata OH-ryhmän sokerikomponentin 3’-päässä, johon seuraavan nukleotidin pitäisi liittyä.

Jos reaktioseokseen lisätään dideoksiATP:a, ilmenee reaktion pysähtyminen templaatissa vastakkaisella puolella tymiinissä (T), kuten kuvassa 18(3.) on esitetty. Reaktio ei kuitenkaan pääty aina ensimmäisessä T:ssä, sillä seoksessa on mukana myös tavallista dATP:a, joka voi liittyä templaattisäikeeseen dideoksinukleotidin sijasta. dATP:n ja ddATP:n suhde on oltava sellainen, että tavallista säiettä polymerisoituu huomattavan helposti ennen kuin ddATP kiinnittyy. Käytännössä sopiva ddNTP:n osuus on noin 5%.

Säikeen syntetisointi tehtiin aikaisemmin neljänä rinnakkaisena reaktiona, sillä ddATP:n lisäksi myös ddTTP, -GTP ja -CTP on lisättävä seokseen. Tuloksena saadaan käytännössä syntyvän DNA-säikeen jokaiseen kohtaan liittymään dideoksinukleotidi, jolloin syntyy kaiken mittaisia DNA-sekvenssejä.

Seuraavassa vaiheessa erotellaan säikeet siten, että niiden pituudet voidaan mitata. Pilkotut molekyylit leimataan Brownin (2001) mukaan nykyisin automatisoidussa menetelmissä käyttäen fluoresoivaa leimaa. Yleensä fluoresoiva leima lisätään kuvan 19 osoittamalla tavalla dideoksinukleotideihin, jolloin jokaisessa syntetisoidussa molekyylissä on leima 3’-päässään.

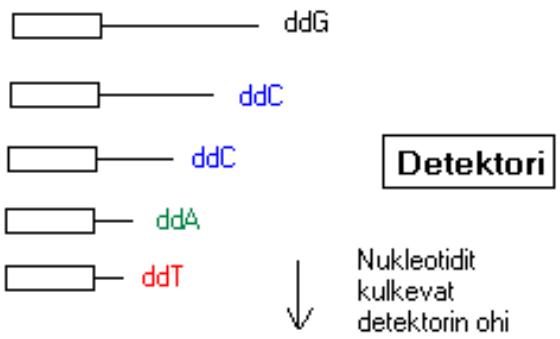
Koska jokaiseen neljään erilaiseen dideoksyNTP:iin on mahdollista liittää erilainen fluoresoiva leima, voidaan neljän rinnakkaisen sekvensointireaktion sijasta käyttää yh-



Kuva 18: DNA:n sekvensointi Sangerin menetelmällä (Brown, 2001).

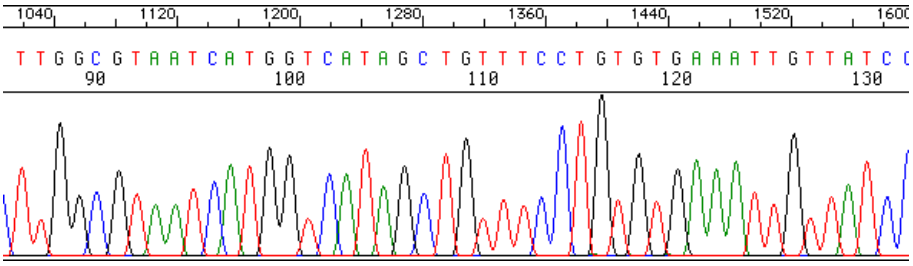
tä ja silti dideoksiNTP:t voidaan erotella toisistaan. Fluoresoivan signaalin luenta tapahtuu automatisoidusti. Ensin näyte laitetaan kapillaarielektroforeesiputkeen, josta se ajetaan kuvan 19 mukaisesti fluoresenssin paljastavan, laseria käyttävän detektorin ohi. Emittoidun signaalin avulla saadaan selville oikea nukleotidi, jonka väri ja sijainti tallennetaan tiedostoon.

Lopulta kun kaikki molekyylit on tutkittu, yhdistää tietokone analyysien antaman tiedon perusteella oikean sekvenssin, joka voidaan tulostaa esimerkiksi kuvan 20 mukaisena elektroferogrammina, jossa jokainen käyrästön huippu vastaa yhtä nukleotidia. Tietokoneohjelma lukee sekvenssin suoraan elektroferogrammista ja voi myös antaa



Kuva 19: Automatisoitu DNA:n sekvensointi (Brown, 2001).

arvion sekvensoinnin luotettavuudesta kunkin nukleotidin osalta. Lopulliset tulokset tallennetaan tekstimuodossa.



Kuva 20: Automatisoidun DNA:n sekvensoinnin tuloste (UMICH, 2004).

Ensimmäiset automatisoidut sekvensointilaitteet tulivat Hunkapillerin (2003) mukaan markkinoille 1980-luvun puolessa välissä. Nykyisillä laitteilla pystytään analysoimaan jo 2 miljoonaa emäsparia vuorokaudessa. Silti tarvitaan vielä paljon tehokkaampia laitteita, jos halutaan todella selvittää esimerkiksi ihmisen genomin sisältämä informaatio. Laitteiden virheherkkyyttä pitäisi pystyä myös vähentämään, jolloin tarvittavia tarkastusajoja voidaan vähentää ja sitäkin kautta nopeuttaa analyysia. Uusista tekniikoista haluttaisiin lisäksi apua kustannusten karsimiseen.

Hillin *& al.* (2000) mukaan on Taq nykyisin käytetyin polymeraasientsyymi. Taq:in etuna on sen toimintakyky korkeassa lämpötilassa, mikä mahdollistaa paremmin GC-pitoisten templaattien käsittelyn. Lämpöstabiilin ominaisuuden vuoksi Taq mahdollistaa tässäkin tutkimuksessa käytetyn, ns. kiertosekvensoinnin (cycle sequencing) käyttämisen. Menetelmässä käytetään korkeaa lämpötilaa denaturoimaan kaksisäikeinen DNA ja se muistuttaa PCR:ää, mutta siinä tarvitaan vain yksi aluke. Etuna on, että näin voidaan sekvensoida useita kierroksia lisäämättä uutta entsyymiä, ja lisäksi tarvitaan

vähemmän templaatti DNA:ta.

Tässä tutkimuksessa sekvensoitu DNA analysoitiin käyttäen BioEdit-ohjelmaa, joka on biologinen sekvenssieditori. Se mahdollistaa perustasolla tapahtuvan proteiini- ja nukleiinihapposekvenssien editoimisen, rinnastuksen ja analysoinnin. BioEdit ei ole erityisen tehokas sekvenssianalyysiohjelma, mutta sen avulla saada tulokset sekä teksti-, graafisessa- (Liite 2.) että matriisimuodossa ja sillä voidaan rinnastaa tutkittavat sekvenssit (Liite 3.).

3.4 Koodaus DNA:lle

Tiedon koodaamisessa DNA:lle oli otettava huomioon kerralla käytettävissä oleva rajoitettu tila sekä toisaalta toistuvien rakenteiden aiheuttamat ongelmat hybridisoitumiselle.

Vaikka DNA-säikeet ovat luonnossa hyvinkin pitkiä, on pitkän, halutun sekvenssin sisältävän säikeen syntetisoiminen vaikeaa ja kallista. Nykyisillä kaupallisilla tekniikoilla pystytään Oligomerin (2004) mukaan syntetisoimaan maksimissaan noin 80:n nukleotidin mittaisia säikeitä, mutta käytännössä paremman tuloksen saa, jos säi on alle 50 nukleotidia pitkä.

Kuvainformaatiota sisältävien sekvenssien luomiseen tutkimuksessa käytettiin valmiiksi syntetisoituja oligonukleotideja, joiden pituudet vaihtelivat 54 - 78 nukleotidin välillä. Kuutta tilattua nukleotidia käytettiin alukkeina luotaessa kolme kuvainformaatiota sisältävää sekvenssiä, joiden pituudet vaihtelivat 88 - 114 nukleotidin välillä. Teoriassa suurin tällä tekniikalla toteutettu sekvenssin pituus on noin 140 nukleotidia, sillä alukkeiden tulee olla komplementaariset toisilleen vähintään 20:n nukleotidin pituudelta.

Sekvenssin alkuun koodattiin kuvan väreistä kertova osuus. Kukin väri koodattiin 14 nukleotidia pitkällä sekvenssillä eli kolmea väriä käytettäessä värien osuus oli 42 nukleotidia. Värien tunnukset oli kuitenkin tehtävä vähintään tämän mittaisiksi, sillä kuvia oli tarkoitus hakea DNA-seoksesta värien komplementin omaavan DNA-säikeen avulla käyttäen hyväksi vastinsäikeiden hybridisoitumista keskenään. Lyhyempien nukeleotidien sitoutuminen templaattiin ei olisi ollut luultavasti tarpeeksi voimakasta erottelemaan templaattteja toisistaan.

Kloonattaessa sekvenssi plasmidiin riitti värialueiden tunnusten lisäksi koodata vain alueiden värin tunnus yhdellä nukleotidilla ja koko kolmella nukleotidilla, sillä plasmideille on olemassa valmiita sekvensointialukkeita. Jos säiettä ei transformoida plasmidiin, mutta se aiotaan sekvensoida, täytyy sen 3’-päähän sisällyttää sekvensointia varten tietty, noin 18 nukleotidia pitkä sekvenssi. Tässä tutkimuksessa käytettiin tiedon tiivistämiseksi 14 nukleotidia pitkää DNA-sekvenssiä apuna sekvensoinnissa.

Koska kolmen värin tunnuksia varten tarvittiin 42 nukleotidia ja sekvensointia varten 14 nukleotidia, jäi kuvan koodaamiseen maksimissaan 84 nukleotidia eli 21 neljän nukleotidin värialuetta. Tämä oli kuitenkin tässä tutkimuksessa riittävä määrä, koska kuvat olivat pieniä ja kuvien yhtenäiset värialueet tarpeeksi suuria.

Jos säiettä ei olisi haluttu ligatoida plasmidiin ja kasvattaa bakteerissa, olisi koodaussysteemiä voitu muuttaa siten, että säikeiden alkuun ei olisi tarvinnut lisätä värien tunnuksia. Tällöin olisi kuvan koodaamiseen saatu 14 nukleotia lisää jokaista väriä kohden.

Tässä tutkimuksessa käytetään värien tunnuksina seuraavia sekvenssejä:

GGGGGGGGGGGGG = keltainen

TTTTTTTTTTTT = musta

AACCAACCAACCAA = punainen

TTCCTTCCTTCCTT = valkoinen

Värien tunnukset on pyritty suunnittelemaan siten, että vain 2-säikeisestä DNA:sta oikean informaation sisältävä säie voi hybridisoitua väriä edustavan alukkeen kanssa. Värien tunnuksena ei voi käyttää toisilleen komplementaarisia sekvenssejä (kuten AAAAAAAAAAAAA ja TTTTTTTTTTTTTT), jotta ”väärä” säie ei hybridisoituisi alukkeen kanssa.

Kutakin yhtenäistä värialuetta edustaa neljän nukleotidin sekvenssi siten, että ensimmäinen nukleotidi kertoo alueen värin ja kolme seuraavaa alueen koon. Koska väri kuvataan vain yhdellä nukleotidilla, voi kuvassa olla maksimissaan neljää eri väriä: G = keltainen, T = musta, C = valkoinen ja A = punainen.

Yhtenäisten värialueiden koko esitetään 3-merkkisenä 4-lukuna: A = 0, T = 1, C = 2 ja G = 3. Kolmesta nukleotidistä vasemmanpuoleinen kertoo lukujen 1 (0 - 3), keskimmäinen lukujen 4 (0 - 3) ja oikeanpuoleinen lukujen 16 (0 - 3) määrän.

Taulukko 1: Lukujen koodaus nelilukuina.

AAA 0	AAT 1	AAC 2	AAG 3	ATA 4	ATT 5	ATC 6	ATG 7
ACA 8	ACT 9	ACC 10	ACG 11	AGA 12	AGT 13	AGC 14	AGG 15
TAA 16	TAT 17	TAC 18	TAG 19	TTA 20	TTT 21	TTC 22	TTG 23
TCA 24	TCT 25	TCC 26	TCG 27	TGA 28	TGT 29	TGC 30	TGG 31
CAA 32	CAT 33	CAC 34	CAG 35	CTA 36	CTT 37	CTC 38	CTG 39
CCA 40	CCT 41	CCC 42	CCG 43	CGA 44	CGT 45	CGC 46	CGG 47
GAA 48	GAT 49	GAC 50	GAG 51	GTA 52	GTT 53	GTC 54	GTG 55
GCA 56	GCT 57	GCC 58	GCG 59	GGA 60	GGT 61	GGC 62	GGG 63

3.5 Alukkeiden suunnittelu PCR-reaktiota varten

Alukkeiden suunnittelu on tärkein vaihe tämän kaltaisessa laboratoriotyössä, sillä ilman hyvin suunniteltuja alukkeita virheiden mahdollisuus kasvaa. Sekä PCR-reaktioissa että sekvensoinnissa tarvittavien alukkeiden suunnittelussa käytettiin apuna *Amplify*-ohjelmaa (Engels 2004). Sen avulla ei voi varsinaisesti suunnitella itse alukkeita, mutta se mahdollistaa valittujen alukkeiden toiminnan testaamisen keskenään tai suhteessa templaattiin.

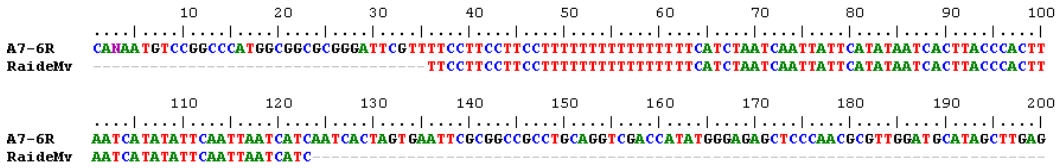
Amplifyn avulla saadaan simuloidun reaktion tuotteiden määrä graafisessa muodossa, eri paksuisina viivoina. Lisäksi ohjelma näyttää alukkeiden sijainnin templaatissa, identtisyyden templaatin kanssa sekä stabiilisuuden. Ohjelma kertoo myös mahdollisista *hairpin*- ja *primer dimer*- rakenteista. Hairpin-rakenteessa DNA-säie hybridisoituu itsensä kanssa muodostaen silmukan. Primer dimer -rakenne tarkoittaa alukkeiden hybridisoitumista keskenään, joka oli tässä tutkimuksessa tarkoituksena syntetisoitaessa kohdassa 3.4 kuvattuja sekvenssejä.

Jos tutkitut alukkeet eivät toimi halutulla tavalla, voidaan Poralin (2001) mukaan, niiden 3’-pään sijaintia tai pituutta muuttamalla usein saada aikaan toimiva ratkaisu, sillä alukkeiden spesifisyys määräytyy etupäässä niiden 3’-pään emäsjärjestyksen mukaan.

PCR- tai sekvensointialukkeiden suunnittelu tietokoneohjelmalla ei takaa reaktion onnistumista, mutta se auttaa poistamaan tai korjaamaan yleisimmät epäonnistumisten syyt, kuten kirjoitusvirheet ja väärin pariutumiset.

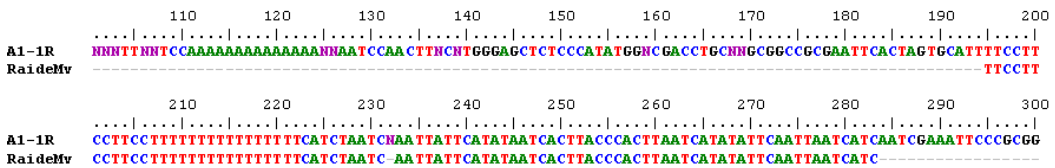
3.6 Tulokset

Parhaiten tutkimus onnistui luvussa 3.3 kuvattua orgaanista muistia käyttäen. *E. coli* -bakteeriin plasmidin mukana transformoitu ja sen jälkeen eristetty DNA oli säilynyt muuttumattomana. Sekvensoinnista saatujen tulosten perusteella voitiin todeta, että molemmat tutkitut bakteeripesäkkeet olivat sisältäneet insertti-DNA:ta.



Kuva 21: Yksityiskohta BioEdit -ohjelman rinnastamista A7-6R -plasmidin DNA:sta ja RaideMV-kuvan sekvenssistä. Kuvasta näkyy, että sekvenssit ovat yhtenevät eli työ on onnistunut.

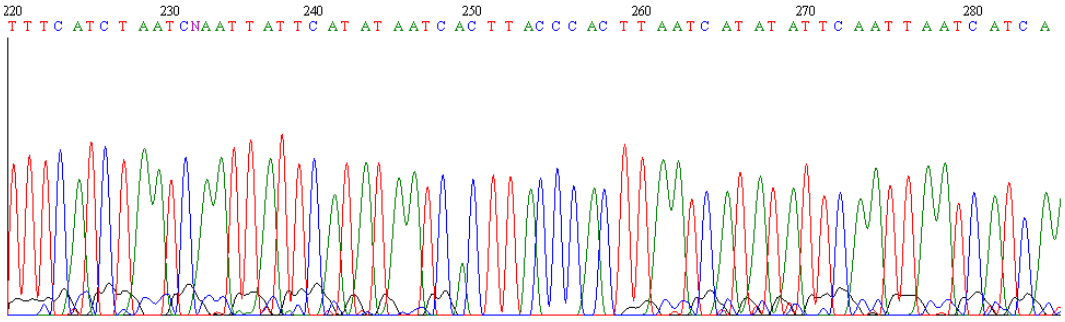
Kuvassa 21 nähdään alkuperäisen RaideMV-kuvan sekvenssi, rinnastettuna BioEdit-ohjelmalla 6R-plasmidista eristetyn ja sekvensoidun DNA:n kanssa. Tämä DNA oli kerätty bakteeripesäkkeestä, jonka ei olisi suurella todennäköisyydellä pitänyt sisältää inserttiä, mutta silti ligaatio oli onnistunut ja plasmidi sisälsi insertin. Kuvasta voidaan nähdä, että sekvenssien rinnastus on onnistunut täydellisesti eli jokaista koodatun RaideMV-kuvan sekvenssin emästä vastaa komplementaarinen emäs sekvensoidussa insertissä. Koko plasmidista sekvensoitu DNA on nähtävissä liitteessä 2, kuvassa 2.



Kuva 22: Yksityiskohta BioEdit -ohjelman rinnastamista A1-1R -plasmidin DNA:sta ja RaideMV-kuvan sekvenssistä. Emäksen 232 kohdalla on havaittavissa ylimääräinen N-kirjain sekvensoidussa plasmidin DNA:ssa.

Sekvensointilaitteen antama tekstimuotoinen tuloste ei aina sisällä yksiselitteistä tietoa tutkitusta DNA:sta. Nytkin pesäkkeen 1R-näytteen sekvenssissä on mukana ylimääräinen N, joka tarkoittaa mahdollista tunnistamatonta emästä. Kuvassa 22 nähdään yksityiskohta alkuperäisen RaideMV-kuvan sekvenssin rinnastuksesta BioEdit-ohjelmalla, 1R-plasmidista eristetyn ja sekvensoidun DNA:n kanssa. Kuvasta voidaan nähdä ylimääräinen N-kirjain nukleotin 232 kohdassa. Muuten sekvenssien rinnastus on onnistunut ilman eroavuuksia eli DNA on virheetön.

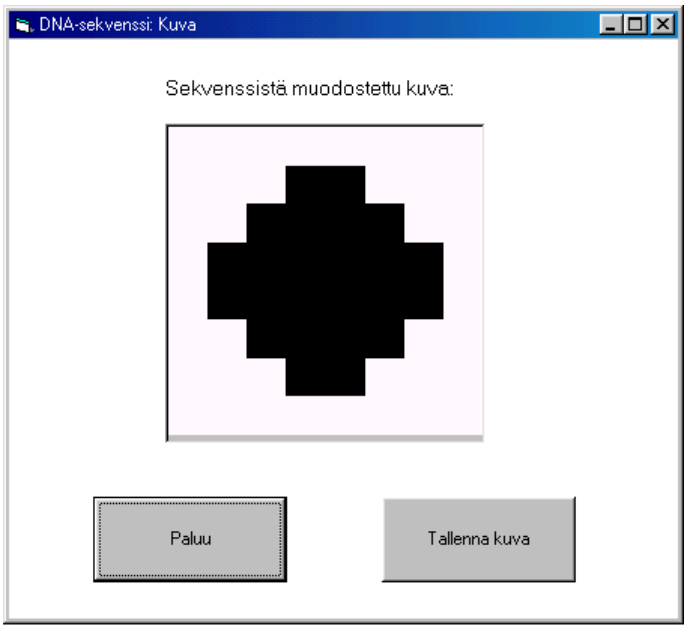
Tutkittaessa tuloksia kuvan 23 mukaisessa graafisessa muodossa, voidaan kuitenkin todeta, että sekvensointi on onnistunut tässäkin tapauksessa hyvin ja N-kirjain tekstimuotoisissa tuloksissa on turha. Kuvasta nähdään selvästi, että N-kirjaimen kohdalla käyrässä ei ole merkittävää huippua, vaan sekvensointilaitteen tekemä tulkinta on väärä.



Kuva 23: Virhe A1-1R -sekvenssissä on nähtävissä emäksen numero 232 kohdalla. Graafisesta esityksestä voidaan huomata, että kyseisellä kohdalla ei ole näkyvissä käyrässä huippua, joten virhe johtuu sekvensointilaitteen väärästä tulkinnasta.

Lukuisista tarkoista työvaiheista ja usean päivän työstä huolimatta kuvan insertointi *E. coli* -bakteeriin ja eristäminen bakteerista onnistui erinomaisesti. Kuvassa 24 on esitettynä DNA-sekvenssi -ohjelman pesäkkeestä 6R eristetyn DNA:n mukaan piirtämä kuva. Koska plasmidista eristetty ja sekvensoitu DNA oli identtinen alkuperäisen sekvenssin kanssa, on myös sekvenssistä muodostettu kuva identtinen alkuperäisen kuvan kanssa.

DNA-sekvenssi -ohjelma ei pystynyt piirtämään kuvaa pesäkkeestä 1R eristetystä DNA:sta, sillä ohjelma ei osaa tulkita mahdollisia virheitä sekvenssissä, koska ohjelma lukee vain tekstimuotoista tietoa. Toisaalta DNA:n sekvensointiohjelmat ja myös useat muut DNA-sekvenssien analysointiohjelmat ovat kykenemättömiä tekemään varmoja päätelmiä saadun tiedon paikkansapitävyydestä ja tarvitsevat usein ihmistä tulkitse-



Kuva 24: DNA-sekvenssi -ohjelman piirtämä kuva plasmidin A7-6R sekvensoidusta DNA:sta. Kuva on identtinen alkuperäisen kuvan kanssa.

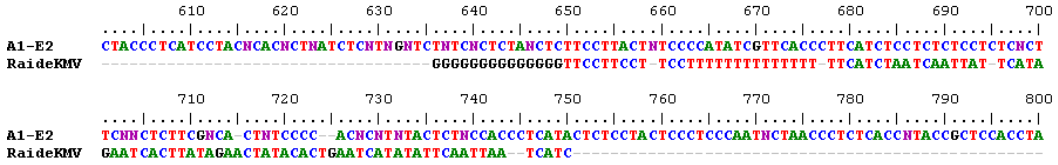
maan tuloksia.

Kohdassa 3.3 kuvattu työn toinen vaihe, jossa yritettiin hakea tietyn värin sisältäviä kuvasekvenssejä liukoisessa muodossa olevasta DNA-tietokannasta epäonnistui pahoin. Liitteessä 2, kuvissa 3 ja 4 on esitettynä keltaisen värin etsinnässä löydetty DNA-sekvenssi. Tavoitteena oli hakea 124 nukleotidia pitkä sekvenssi, mutta tulos oli 1015:n nukleotidin mittainen. Liitteessä 3, kuvassa 4 on rinnastettuna sekvensoitu DNA ja alkuperäisen kuvan sekvenssi. Vaikka sekvensseissä voidaankin havaita tiettyjä yhtäläisyyksiä, ovat ne kuitenkin hyvin erilaisia.

Vertailtaessa sekvenssejä kokonaisuudessaan liitteen 3, kuvan 3 mukaisesti, voidaan todeta, että sekvensoitu DNA sisältää paljon osia, jotka voivat olla lähtöisin myös muista kuvista. Tämä tukisi päätelmää, että värin mukainen tietokantahaku ei ole onnistunut, vaan mukaan on tullut myös ei-toivottuja sekvenssejä, jotka ovat lopulta sotkeneet sekvensointireaktion. Haluttua paljon pidempi DNA voi olla tuloksena väärien sekvenssien sekoittumisesta reaktioon.

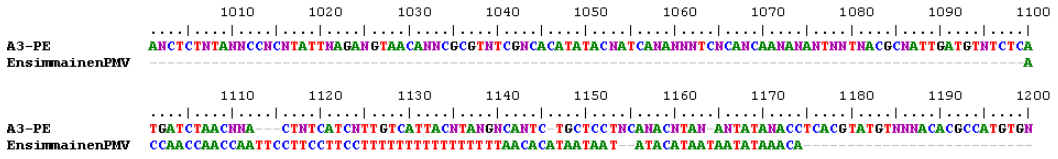
Mahdollista on kuitenkin myös, että sekvensointialuke on ollut huonosti suunniteltu. Vaikka alukkeen suunnittelussa käytettiin apuna Amplify-ohjelmaa, on silti mahdollista, kuten kohdassa 3.5 todettiin, että simuloitu reaktio ei vastaa todellista tilannetta ja

siksi sekvensointi ei onnistu halutulla tavalla. Jos aluke on pystynyt hybridisoitumaan useampaan kuin yhteen kohtaan halutussa sekvenssissä, saadaan tuloksena luultavasti liian pitkä sekvenssi. Tämän sekvenssin pitäisi kuitenkin sisältää paljolti samoja osia alkuperäisestä kuvasta, eikä suinkaan osia eri kuvista.



Kuva 25: Yksityiskohta BioEdit-ohjelman rinnastamista A1-E2 -DNA:sta ja RaideKMV-kuvan sekvenssistä. Kuvasta voi nähdä, että sekvensoitu A1-E2 poikkeaa huomattavasti halutusta.

Liitteessä 2, kuvissa 5 ja 6 on esitettynä punaisen värin etsinnässä löydetty DNA-sekvenssi. Tavoitteena oli saada haettua tietokannasta 88 nukleotidia pitkä sekvenssi, mutta tulos oli 1366:n nukleotidin mittainen. Kuvassa 26 on rinnastettuna sekvensoitu DNA ja alkuperäisen kuvan sekvenssi. Vaikka sekvensseissä voidaankin havaita samankaltaisia rakenteita, ei rinnastus anna kuitenkaan kunnollista tulosta. Tulos on siinä suhteessa vertailukelpoinen keltaisella värillä suoritetun kalastuksen kanssa, että myöskään tässä tapauksessa ei voida varmasti osittaa epäonnistumisen syytä.



Kuva 26: Yksityiskohta BioEdit-ohjelman rinnastamista A3-PE -DNA:sta ja EnsimmainenPMV-kuvan sekvenssistä. Kuvasta voi nähdä, että sekvensoitu A3-PE poikkeaa huomattavasti halutusta.

Tässä luvussa kerrottiin miten tutkimus toteutettiin käytännössä. Aluksi kuvattiin lyhyesti tutkimuksen taustaa ja sen vaiheet, seuraavaksi kerrottiin DNA-sekvenssi -ohjelman tarkoitus ja toiminnot. Lisäksi luvussa kuvattiin tutkimuksen molekyylibiologian osuus, DNA:lle tapahtunut koodauksen toteutus sekä alukkeiden suunnittelu. Luvun lopuksi kerrottiin tutkimuksen tulokset.

Seuraavassa luvussa pohditaan DNA-laskennan ja DNA-muistin käyttämiseen liittyviä yleisiä ongelmia. Lisäksi käydään läpi tutkimukseen liittyneitä ongelmallisia tilanteita

ja virheiden mahdollisia syitä, sekä esitetään vaihtoehtoisia tapoja ratkaista hankaluuksia aiheuttaneet tutkimuksen osat.

4 Pohdinta

Tämän luvun kohdassa 4.1 pohditaan aluksi DNA-laskennan ja DNA-muistin käyttämiseen liittyviä yleisiä ongelmia. Kohdassa 4.2 käydään läpi tutkimukseen liittyneitä ongelmallisia tilanteita sekä esitetään vaihtoehtoisia tapoja ratkaista hankaluukisia aiheuttaneet tutkimuksen osat. Kohdassa 4.3 pohditaan vielä erikseen sekvensointiin liittyviä virhemahdollisuuksia.

4.1 Pohdintaa DNA-laskentaan ja -muistiin liittyen

Vaikka DNA-laskennan avulla kyetään ratkaisemaan NP-täydellisiä ongelmia, soveltuvat DNA-tietokoneet luultavasti paremmin ratkaisemaan muita, erilaisia ongelma-alueita. DNA kyllä mahdollistaa rinnakkaisen reaktioiden käyttämisen, mutta usein esivalmistelut sekä reaktiot vievät paljon aikaa. Myös tulokset voivat olla hyvästä suunnittelusta ja toteutuksesta huolimatta tulkinnanvaraisia. Enää ei uskotakaan, että DNA-tietokoneet voivat syrjäyttää perinteisiä tietokoneita, mutta molekylaarisen laskennan perustutkimukseen ja sovelluksiin tulisi suunnata resursseja. Nykyisin DNA-laskennan tutkimus onkin suuntautumassa matemaattisesta laskennasta kohti laajempia nano- ja bioteknologian sovelluksia, sillä usko DNA:ssa piileviin suuriin mahdollisuuksiin on yhä olemassa.

DNA:n käyttäminen orgaanisena muistina on hyvin houkutteleva ajatus, sillä tarjoaahan se periaatteessa nykyisiin menetelmiin verrattuna monituhatkertaisen tallennuskapasiteetin. Toisaalta DNA:lle tapahtuva tiedontallennus on nykyisin vielä vaikeaa toteuttaa ja se on myöskin hyvin kallista. Vaikka tieto voidaan pakata DNA:lle hyvin pieneen tilaa, on suurien tiedostojen luominen hankalaa, sillä pitkät sekvenssit pitää koota osista. Myös tietokantojen luominen ja tiedonhaku aiheuttaa ongelmia.

Lisäksi DNA:n säilyvyys voi olla ongelmallista, vaikka DNA-sekvenssi voikin pysyä sopivissa olosuhteissa muuttumattomana satojatuhansia vuosia. Kuivattu tai pakastettu DNA säilyy erittäin hyvin, mutta vaatii kuitenkin erityiset säilytystilat. Bakteerien käyttäminen tiedon tallennukseen on puolestaan mielestäni kohtuullisen hankalasti toteutettavissa oleva vaihtoehto, kuten esitin luvussa 2.3.

Esimerkiksi Harlanin *& al.* (2003) tutkimuksessa ei kerrottu keinoja, joiden avulla bakteerikolonisaatioihin tallennettu tieto voitaisiin saada helposti käyttöön. Tiedon helppo

saatavuus on kuitenkin tärkeää, sillä muuten DNA-muisti ei voi kilpailla perinteisten ratkaisujen kanssa. Nykyiset tekniikat eivät myöskään mahdollista sekvenssien löytämistä bakteerien DNA:sta ilman jokaiselle tietopaketille yksilöllistä tunnistussekvenssiä. Ja vaikka bakteereihin tallennettuihin sekvensseihin koodattaisiinkin yksilöllinen tunnistussekvenssi, on tiedon etsiminen kohtalaisen hidasta ja vaikeaa.

Ongelmana on myös se, että vaikka käytetty bakteeri kestäisikin äärimmäisiä olosuhteita, ei voida olettaa, että se säilyy hengissä ilman valvottuja olosuhteita. Lisäksi bakteerin kasvatuspaikan on oltava sellainen, että bakteeri on helposti saatavilla. Toisaalta on vaikeaa taata, että jokainen miljooniin bakteereihin tallennettu tiedon osa pysyy tallessa, jos eri tietoa sisältävät bakteerit ovat samassa tilassa.

DNA-sirut ovat suureksi hyödyksi monissa biologisissa tutkimuksissa, ja ne mahdollistavatkin DNA:lle tallennetun tiedon tulkitsemisen laajassa mittakaavassa, mutta niiden käyttäminen tietokantana on hankalaa niiden lyhyen säilyvyyden takia.

Yksi, luultavasti tulevaisuudessa paljon keskustelua herättävä aihe on tiedon tallentamiseen ihmisen omiin soluihin. Nykyisessä asenneilmapiirissä ihmisen DNA:n muokkaaminen on kuitenkin luultavasti vaikeasti hyväksyttävä ajatus. Oma DNA voisi olla kuitenkin varteenotettava paikka tallentaa yksilölle tärkeää tietoa. Etuna olisi ainakin se, että tieto kulkisi aina huomaamattomasti mukana ja muiden olisi vaikeaa käyttää sitä hyväksi ilman lupaa.

4.2 Pohdinta tutkimuksen osalta

Yksi merkittävä ongelma tässä tutkimuksessa oli liian vähäinen käytettyjen värien määrä. Värejä olisi pitänyt ottaa mukaan alunperin ainakin kahdeksan. DNA-sekvenssi -ohjelmalle värien määrä ei olisi ollut mikään ongelma, mutta DNA:lle koodauksen suhteen tutkimus olisi hankaloitunut hieman, sillä sekvenssit olisi pitänyt suunnitella pidemmiksi. Vaikka kustannukset olisivat nousseet ja tutkimus vaikeutunut, olisi värien lisääminen kuitenkin kannattanut. Jos käytössä on vain neljä väriä, ei voida tehdä kolmea kolmiväristä kuvaa, yhdistää DNA:ta ja suorittaa hakua kolmella eri värillä, sillä vähintään kahdessa kuvassa on käytössä sama väri, eikä käytössä ole menetelmää erotella samanvärisiä sekvenssejä toisistaan. Lopulta asialla ei kuitenkaan olisi ollut suurempaa merkitystä, sillä tietokantahaut värien avulla epäonnistuivat pienemmälläkin värien määrällä. Toisaalta jos värejä olisi ollut käytössä useampi, olisi *E. coli*

-bakteeriin voitu transformoida näyttävämpi kuva, sillä nyt tallennettu kuva oli hyvin yksinkertainen.

Ongelmana DNA-laskennan menetelmissä on käytettävissä olevien operaattorien määrä, sillä käytännössä seoksesta voidaan etsiä tiettyä sekvenssiä käyttämällä ”=”-operaattoria. Jos halutaan karsia tietyt sekvenssit pois, voidaan ne kalastaa seoksesta eli käyttää ”≠”-operaattoria, mutta edelleen ongelmana on erotella jäljelle jääneet DNA-sekvenssit. ”≤”- ja ”≥”-operaattorien käyttäminen on käytännössä erittäin vaikeaa toteuttaa koeputkessa.

Kaikkia kuvia ei voida tehdä nykytekniikalla, ainakaan käyttämällä tämän tutkimuksen metodia DNA-sekvenssien luomiseen, sillä usein toistuvat sekvenssit ovat ongelma PCR:ssä. Esimerkiksi rakenne CAAT TAAT CAAT TAAT, jossa kuvassa toistuu valkoinen ja musta bikseli useampaan kertaan, saisi luultavasti PCR:n tuottamaan jotain täysin halutusta poikkeavaa.

Jos sekvenssit halutaan toteuttaa kohdassa 2.4 kuvatulla oikeaoppisella tavalla, kasvaa sekvenssien pituus niin pitkäksi, että tällaisen tutkimuksen toteuttaminen käy mahdottomaaksi nykyisin käytössä olevilla märkälaboratoriomenetelmillä. Plasmideihin voidaan kyllä ligatoida pitkiä sekvenssejä, mutta niiden syntetisoiminen on paljon vaikeampaa.

Sekvenssien tietokantahaku, jossa ei käytetty plasmidia sekvenssin tallentamiseen, aiheutti sen, että sekvenssin loppuun piti lisätä yhteinen sekvenssi sekvensointi-PCR:ää varten. Tämä puolestaan kasvatti käytetyn säikeen pituutta 14:llä nukleotidilla, joka on suhteellisen suuri osa käytettävissä olevasta tilasta.

Myös värien etsintä alukkeen avulla vaatii noin 18 nukleotidia pitkän koettimen käyttämistä. Tämä tarkoittaa sitä, että jokaista kuvan väriä varten on sekvenssiin lisättävä vastaavan mittainen, alukkeelle komplementaarinen sekvenssi eli värien määrän kasvaessa, kasvaa myös sekvenssin pituus huomattavasti. Molekyylibiologisten menetelmien kehittyminen tulee kuitenkin luultavasti poistamaan syntetisoitavan sekvenssin pituuteen liittyvät ongelmat muutaman vuoden sisällä, jolloin tiedon koodaamiseen käytettävissä oleva tila myöskin kasvaa.

4.3 Mahdolliset virheet sekvensoinnissa

Hillin *& al.* (2000) mukaan sekvensointituloksen laatuun negatiivisesti vaikuttavia seikkoja voivat olla kontaminantit tai muut epäoptimaaliset seikat sekvensoinnin aikana. Yleisimmin ongelmia ilmenee sekvensointireaktion aikana tai sen jälkeisessä prosessoinnissa. Huonon sekvensointituloksen saa helposti jos käytössä on epäpuhdas templaatti, templaatti- tai alukepitoisuus on väärä, alukkeen sitoutuminen ei onnistu optimaalisesti tai värjättyjen sitoutumattomien dideoksinukleotidien poistaminen ei ole tarpeeksi tehokasta.

Tässä tutkimuksessa pyrittiin tekemään kaikki työvaiheet mahdollisimman huolellisesti, mutta virheiden mahdollisuus on aina olemassa. Lisäksi haettaessa DNA-sekvenssejä biotiinilla leimattujen alukkeiden avulla, jouduttiin käytössä olleita ohjeita soveltamaan hyvinkin radikaalisti, sillä käytettävissä ei ollut tutkimukseen suoraan käyttökelpoista ohjetta. Näistä seikoista johtuen sekvensoitavan DNA:n sekaan on voinut jäädä epäpuhtauksia, jotka ovat vääristäneet tuloksia.

Hyvän sekvenssisignaalin saaminen edellyttää myös tarpeeksi suuren templaatti-alukepolymeraasi -kompleksimäärän syntymistä. Siihen puolestaan vaikuttaa voimakkaimmin alukkeen sitoutuminen, joten alukkeiden riittävään määrään ja spesifiseen suunnitteluun tulee kiinnittää huomiota. Tutkimuksessa käytetyissä sekvensseissä oli suuria eroavaisuuksia GC-pitoisuuksien suhteen, joten säikeiden sulamislämpötiloissa saattoi olla huomattavia eroja. Sulamislämpötilojen ero on voinut vaikuttaa sekvensointireaktioon ja sitä kautta heikentää tutkimuksen tuloksia.

Myös templaattien määrän tulee olla sopiva, sillä niiden vähäinen määrä heikentää signaalia ja lisää taustakohinaa. Liian korkea templaattipitoisuus aiheuttaa alussa erittäin voimakkaan, mutta nopeasti heikkenevän signaalin. Templaattien laadun vaihtelu voi johtua useista eri seikoista, kuten esimerkiksi alkuperäisestä viljelmästä tai plasmidin isäntänä käytetystä bakteerista. Tutkimuksessa templaattina käytetyn sekvensoitavan DNA:n pitoisuutta oli vaikeaa määrittää tarkasti, joten väärä pitoisuus on voinut myös vaikuttaa heikentävästi sekvensoinnin tulokseen.

Korkea nukleaasikonsentraatio on haitallisimmillaan jos käytetään radioaktiivisia metodeja, mutta se voi häiritä myös käytettäessä fluoresoivia leimoja, kuten tässä tutkimuksessa tehtiin. RNA-kontaminointi on harvinaista, mutta templaattina käytetyn DNA:n käsittely RNase:lla hajottaa mahdollisen RNA:n. Suolan aiheuttamat kontami-

naatiot huonontavat sekvensoinnin laatua inhiboimalla reaktioita, joten huolella tehdyt saostukset ja niiden jälkeiset pesut vähentävät virheiden riskiä. Tässä tutkimuksessa oli ajoittain ongelmia pesujen suhteen, käytettäessä magneettia Dynabeadien erotteluun, mutta ongelmat koskivat lähinnä potentiaalista näytteiden häviämistä, eikä niinkään epäpuhtauksien jäämistä näytteeseen.

Tässä luvussa pohdittiin DNA-laskennan ja DNA-muistin käyttämiseen sekä sekvensointiin liittyviä yleisiä ongelmia. Lisäksi käytiin läpi tutkimukseen liittyneitä ongelmallisia tilanteita sekä esitettiin vaihtoehtoisia tapoja ratkaista hankaluuksia aiheuttaneet tutkimuksen osat. Seuraavassa luvussa esitetään lyhyt yhteenveto tutkimuksen keskeisistä tuloksista.

5 Yhteenveto

Nykyiset molekyylibiologian työmenetelmien kehittymättömyys rajoittaa huomattavasti erilaisia sovelluksia, joissa voidaan käyttää DNA:ta tiedon tallentamiseen. DNA:ta on pyritty käyttämään apuna ratkaistaessa laskennallisia ongelmia jo yli kymmenen vuoden ajan, mutta vieläkin ollaan kaukana todellisista käytännön sovelluksista. DNA-tietokoneet eivät luultavasti tulekaan koskaan korvaamaan perinteisiä tietokoneita, mutta ehkä DNA:ta voidaan käyttää apuna ainakin tiedon tallennnuksessa. Eräs nykyisin tutkittu sovellusalue on DNA-muisti, jossa käytetään bakteerin DNA:ta tiedon tallentamiseen, mutta syntetisoitua DNA:ta voidaan periaateessa lisätä mihin tahansa soluihin.

Nykyisellään DNA:n käsittely on kuitenkin kallista. Lisäksi esivalmistelut, varsinainen työ laboratoriossa sekä reaktiot vievät paljon aikaa. Tämän lisäksi tulokset voivat olla hyvästä suunnittelusta ja toteutuksesta huolimatta tulkinnanvaraisia.

Tässä tutkimuksessa kuvattiin tapa, jolla voitiin nykyisiä menetelmiä käyttäen tallentaa tietoa DNA:lle. Tutkimuksessa käytettäväksi valmistettiin tietokoneohjelma, jolla pystyttiin muuttamaan kuvia DNA-koodeiksi ja DNA-sekvenssejä kuviksi. Tutkimuksen molekyylibiologian osuus koostui kahdesta erilaisesta tavasta käyttää DNA:ta tiedon tallentamiseen. Osuus, jossa *E. coli* -bakteerin plasmidiseen DNA:han tallennettiin koodattuna pieni bittikartta -kuva osoitti, että bakteereita voidaan todella käyttää tiedon tallentamiseen.

Toisessa tutkimuksen osuudessa DNA-tietokannasta yritettiin erotella kuvia värien mukaan, käyttäen apuna biotiinilla leimattuja koettimia. Tietokantahaut osoittautuivat kuitenkin vaikeiksi toteuttaa. Jos halutaan käyttää laajaa tietokantaa, ovat DNA-sirut luultavasti tämän hetkisistä sovelluksista käyttökelpoisin vaihtoehto. Ne ovat suureksi hyödyksi monissa biologisissa tutkimuksissa, mutta niiden käyttäminen pitkäaikaisena tietokantana on mahdotonta niiden lyhyen säilyvyyden takia.

DNA:n käyttäminen muistina tarjoaa periaatteessa nykyisiin menetelmiin verrattuna monituhatkertaisen tallennuskapasiteetin. Toisaalta vaikka tieto voidaankin pakata DNA:lle hyvin pieneen tilaa, on suurien tiedostojen luominen hankalaa, sillä pitkät sekvenssit pitää koota osista.

Kaikkea tietoa ei voida myöskään koodata helposti nykytekniikalla, sillä toistuvat se-

kvenssit ovat ongelma PCR:ssä. Jos sekvenssit halutaan toteuttaa oikeaoppisesti koodaten, kasvaa sekvenssien pituus niin paljon, että tietokannan toteuttaminen käy mahdottomaksi nykyisin käytössä olevilla märkälaboratoriomenetelmillä. Plasmideihin voidaan nykyisin ligatoida pitkiäkin sekvenssejä, mutta niiden syntetisoiminen on paljon vaikeampaa.

Myös DNA tietokannan säilyttäminen on hankalaa, vaikka DNA-sekvenssi voikin sopivissa olosuhteissa säilyä hyvinkin pitkiä aikoja. Kuivattu tai pakastettu DNA säilyy erittäin hyvin, mutta vaatii kuitenkin erityiset säilytystilat. Bakteerien käyttäminen tiedon tallennukseen ja etenkin halutun tiedon hakeminen bakteerien plasmideista on kohtuullisen hankalaa. Tiedon helppo saatavuus on kuitenkin tärkeää, sillä muuten DNA-muisti ei voi kilpailla perinteisten ratkaisujen kanssa. Ja vaikka käytetyt bakteerit kestäisivätkin äärimmäisiä olosuhteita, ei voida olettaa, että ne säilyvät vahingoittumattomina valvomattomissa olosuhteissa.

Luultavasti tulevaisuudessa paljon keskustelua herättävä, mutta varteenotettava vaihtoehto on tiedon tallentamiseen ihmisen omiin soluihin. Oma DNA voisi olla oivallinen paikka tallentaa yksilölle tärkeää tietoa, sillä siten tieto kulkisi aina mukana ja olisi vaikeasti väärinkäytettävissä.

Viitteet

Adleman, L.M. (1998) Computing with DNA. *Scientific American* **2**(279), 54-61.

Arita, M. (2004) Writing Information Into DNA *Aspects of Molecular Computing* (toim. Jonoska, N., Paun, G., Rozenberg, G.), Springer Verlag, New York, LNCS(2950), 23-35.

Arita, M., Kobayashi S. (2002) DNA Sequence Design Using Templates *New Generation Computing* **20**(3), 263-277.

Berg, J.M., Tymoczko J.L., Stryer L. (2002) *Biochemistry*. W.H. Freeman and Company, New York.

Braich, R.S., Chelyapov N., Jonson C., Rothemund P., Adleman L. (2002) 3-SAT Problem on a DNA Computer. *Science* **296**, 499-502.

Brown, T.A. (2001) *Gene Cloning and DNA analysis*. Blackwell Science Ltd, Oxford.

Babcock, D., Bochannek, A. (2004) *The Computer History Museum*. WWW-sivusto, `http://www.computerhistory.org/` (16.3.2005).

Deaton, R., Chen, J., Bi, H., Rose, J., A. (2003) A Software Tool for Generating Noncrosshybridizing Libraries of DNA Oligonucleotides *DNA Computing:8th International Workshop on DNA-Based Computers* (toim. Hagiya, M., Ohuchi, A.), Springer LNCS, Heidelberg, 252 - 261.

Engels, B (2001) *Amplify*. WWW-sivusto, `http://engels.genetics.wisc.edu/amplify/` (20.9.2004).

Fagerholm, J. (2004) Dataähky. *Tietoyhteys* **4/2004**, 17.

Hakala, M. (2004) Piioksidin korvaava eristemateriaali. *Tietoyhteys* **3/2004**, 30.

Hagiya, M. (2004) Towards Molecular Programming - a Personal Report on DNA8 and Molecular Computing *Modelling in Molecular Biology* (toim. Ciobanu, G., Rozenberg, G.), Springer, Berlin, 125-140.

Harlan, F., Wong, P. C., Wong, K.(2003) Organic Data Memory, Using the DNA Approach. *Communications of the ACM* **46**(1), 95-98.

Hill, A., J., M., Helps, N., R. (2000) *A Guide to Automated DNA Sequencing*. (Saatavana myös: http://www.dnaseq.co.uk/Downloads/sequencing_guide.pdf, 24.09.2004).

Oligomer Oy (2004) *Oligomer*. WWW-sivusto, http://www.oligomer.fi/ (20.9.2004).

Pâun, G., Rozenberg, G., Salomaa, A. (1998) *DNA Computing*. Springer–Verlag, Heidelberg.

Porali, I. (2001) *PCR:ää biteillä: Amplify*. http://www.csc.fi/lehdet/atcsc/atcsc5-98/amplify.html (20.9.2004).

Sakamoto, K., Gousa, H., Yokoyama, S., Yokomori, T., Hagiya, M. (2000) Molecular Computation by DNA Hairpin Formation. *Science* **288**, 1223-1226.

Sipser, M. (1997) *Introduction to the Theory of Computing*. PWS Publishing Company, Boston.

Suomen bioteollisuus (2003) *Suomen Bioteollisuuden (FIB) biotekniikan sanasto*. WWW-sivusto, http://www.finbio.net/sanasto/ (1.10.2004)

Tulpan, D., Hoos H., Condon A. (2003) Stochastic local search algorithms for DNA word design *DNA Computing:8th International Workshop on DNA-Based Computers* (toim. Hagiya, M., Ohuchi, A.), Springer LNCS, Heidelberg, 229-241.

UMICH (2004) *How do we Sequence DNA?*. WWW-sivusto, http://seqcore.brcf.med.umich.edu/doc/educ/dnapr/sequencing.html (10.9.2004).

Webopedia (2004) *Online Computer Dictionary for Computer and Internet Terms and Definitions*. WWW-sivusto, http://www.webopedia.com/ (16.3.2005).

Wong G. (2003) DNA Microarray Data Analysis *DNA Microarray Data Analysis* (toim. Tuimala, J., Laine, M., M.), CSC-Scientific Computing Ltd., Helsinki, 15 - 24.

Liite 1: Tutkimuksen keskeiset termit

Tutkimuksen keskeinen käsite on orgaaninen DNA-muisti. Tutkimuksessa tallennettiin *E. coli* -bakteeriin, käyttäen plasmidin DNA:ta, pienen bittikartta -kuvan koodi. Tutkimuksen toinen keskeinen käsite on DNA-laskenta, sillä vaikka tutkimuksessa ei käytettykään DNA:ta apuna laskennassa, yritettiin tutkimuksessa käyttää DNA-laskennan menetelmiä etsittäessä tietyt kriteerit täyttäviä DNA-sekvenssejä DNA-seoksesta.

Lisäksi tämän tutkimuksen aihealueeseen liittyvät keskeisesti seuraavat termit.

Aluke eli primeeri on lyhyt polynukleotidiketju, jota käytetään käynnistämään DNA:n kahdentuminen (Heikkinen *& al.*, 2002).

cDNA-kirjasto on kokoelma plasmidiin pakattuja komplementaarisia DNA-säikeitä, jotka on syntetisoitu käyttäen mallina RNA:ta (Heikkinen *& al.*, 2002).

Denaturaatio on tapahtuma, jossa kaksisäikeisen DNA:n säikeet voidaan erottaa toisistaan käyttämällä joko happoa tai emästä, tai lämmittämällä liuos n. $85\,^\circ - 95\,^\circ\mathrm{C}$:n lämpötilaan. Denaturointi on helppoa toteuttaa, koska emäsparien väliset vetysidokset ovat hyvin heikkoja (Berg *& al.*, 2002).

Deoksinukleotidit ovat 2’-deoksiadenosiini 5’-trifosfaatti (dATP), 2’-deoksitymiini 5’-trifosfaatti (dTTP), 2’-deoksiguanosiini 5’-trifosfaatti (dGTP), 2’-deoksisytidiini 5’-trifosfaatti (dCTP) (Heikkinen *& al.*, 2002).

Dideoksinukleotidi on nukleotidi, jossa deoksiriboosi on korvattu dideoksiriboosilla, johon DNA-polymeraasit eivät pysty liittämään uusia nukleotideja. Dideoksinukleotidin lisääminen säikeen päähän johtaa polymerisaation pysähtymiseen (Heikkinen *& al.*, 2002).

DNA tarkoittaa deoksiribonukleiinihappoa. Se muodostuu sokeri-fosfaatti -rungosta sekä siihen liittyvistä, vaihtuvista emäksistä adeniini (A), guaniini (G), tymiini (T) ja sytosiini (C). (Berg *& al.*, 2002).

DNA-laskenta tarkoittaa laskentaa DNA-molekyylien avulla. Kärjistetysti esitettynä DNA-laskennassa on ideana siirtyminen piistä hiileen ja mikropiireistä DNA-molekyyleihin. Tarkoituksena on käyttää hyväksi orgaanisten molekyylien kykyä käsitellä tietoa ja siten osittain korvata perinteiset tietokoneet (Hagiya, 2003).

DNA-siru (DNA microarray) on alustalle kiinnitetty kokoelma DNA-molekyylejä, jotka voidaan hybridisoida liuoksessa olevien nukleiinihappojen kanssa (Heikkinen *& al.*, 2002).

E. coli (Kolibakteeri, Escherichia coli) on suolistobakteeri, jota käytetään geenitekniikassa yleisesti isäntä- eli tuotantosoluna (Suomen Bioteollisuus, 2003).

Elektroporaatio tarkoittaa sähköimpulssien käyttöä avaamaan solukalvon siten, että plasmidi saadaan siirrettyä bakteerin sisälle (Heikkinen *& al.*, 2002).

Emäspari (bp, base pair) tarkoittaa nukleotidien emäksiä, jotka kaksisäikeisessä DNA:ssa sitoutuvat toisiinsa vetysidoksilla. DNA:ssa on 4 emästä: adeniini (A), guaniini (G), tymiini (T) ja sytosiini (C). Emäspareilla tarkoitetaan energeettisesti edullisimpia pareja A-T ja C-G (Heikkinen *& al.*, 2002).

Formaali kieli on kieli, jolla on ennalta määritelty kielioppi, mutta jota ei ole tarkoitettu puhuttavaksi (Sipser, 1997).

Genomi (perimä, genome) tarkoittaa eliön tai solun kromosomien sisältämää perinnöllistä informaatiota (Heikkinen *& al.*, 2002).

Hybridisaatio (annealing, renaturaatio) on tapahtuma, jossa laskemalla lämpötilaa saadaan komplementaariset yksisäikeiset DNA-molekyylit liittymään yhteen (Heikkinen *& al.*, 2002).

Insertti on DNA:n osa, joka liitetään vektoriin esim. kloonausta varten (Suomen Bioteollisuus, 2003).

In vitro tarkoittaa toimenpidettä, joka tehdään keinotekoisissa olosuhteissa esimerkiksi koeputkessa tai kasvatusalustalla (Heikkinen *& al.*, 2002).

Kilobase (kb, kbp, kiloemäspari) tarkoittaa tuhannen nukleotidin mittaista DNA- tai RNA-jaksoa (Heikkinen *& al.*, 2002).

Kloonaus on geenin monistamista. Tässä tutkimuksessa kloonauksella tarkoitetaan plasmidin ja siihen insertoidun DNA:n monistamista *E. coli* -bakteerissa (Heikkinen *& al.*, 2002).

Koetin on radioaktiivisesti tai kemiallisesti leimattu DNA- tai RNA-jakso, jota käytetään hybridisaatiossa tietyn nukleiinihapon etsimiseen. Koetin on yleensä vähintään

20:n emäksen pituinen tunnettu DNA-jakso. (Suomen Bioteollisuus, 2003).

Kohesiivinen pää tarkoittaa kaksisäikeistä DNA:ta, jonka toinen säie on toista pidempi (Heikkinen *& al.*, 2002).

Komplementaarinen (complementary) tarkoittaa nukleiinihapposäiettä, jonka jokainen nukleotidiemäs pariutuu Watson-Crick -periaatteen mukaisesti toisen säikeen kanssa (Heikkinen *& al.*, 2002).

Kromosomi tarkoittaa DNA-säiettä, joka aitotumallisilla pakkautuu tiiviisti histoniproteiinien ansiosta. Pääosa solun geenistöstä on kromosomeissa. (Heikkinen *& al.*, 2002).

Ligaasi on entsyymi, joka kykenee yhdistämään katkaistuja DNA-säikeitä (Berg *& al.*, 2002).

Mooren laki on Gordon Mooren vuonna 1965 luoma laki mikropiirien komponettien määrän kasvusta. Alunperin Moore esitti mikropiireissä käytettyjen komponenttien määrän kaksinkertaistuvan 12 kuukauden välein ja päätteli, että trendi jatkuisi pitkälle tulevaisuuteen. Myöhemmin vauhti on hidastunut hieman, mutta mikropiirien tallennuskapasiteetti kaksinkertaistuu silti 18 kuukauden välein. Tämä onkin Mooren lain nykyinen määritelmä ja sen uskotaan pitävän paikkansa ainakin kaksi seuraavaa vuosikymmnentä (Webopedia, 2004).

Mutaatio tarkoittaa organismin DNA-emäsjärjestyksen muuttumista ilman rekombinaatiota (Heikkinen *& al.*, 2002).

NP-täydellinen ongelma on sellainen laskennallinen ongelma, jolle ei ole löydetty polynomisessa ajassa toimivaa determinististä ratkaisua. Käytännössä tämä tarkoittaa sitä, että kyseisiä ongelmia ei pystytä ratkaisemaan nykyisillä tietokoneilla, sillä suoritusaika kasvaa eksponentiaalisesti syötteen kokoon verrattuna. Tunnettuja NP-täydellisiä ongelmia ovat mm. Toteutuvuusongelma (SAT), Kauppamatkustajan ongelma (TSP), Hamiltonin polku -ongelma (HP) sekä Solmupeite (VC) (Sipserin, 1997).

Nukleotidi on nukleiinihapon rakenneosa, koostuu puriini- tai pyrimidiiniemäksestä, pentoosisokerista ja fosfaatista (Berg *& al.*, 2002).

Oligonukleotidi tarkoittaa korkeintaan muutamien kymmenten emäksien mittaista DNA-sekvenssiä (Heikkinen *& al.*, 2002).

PCR eli polymeraasiketjureaktio on DNA:n monistamistekniikkaa, jossa saadaan polymeraasientsyymin avulla hetkessä monistettua suuri määrä identtisiä kopioita olemassa olevasta DNA-säikeestä (Berg *& al.*, 2002).

Plasmidi on bakteerien pieni rengasmainen DNA-molekyyli. Käytetään kuljettimena ("geenitaksina"), johon siirrettävä geeni liitetään ja joka kuljettaa geenin vastaanottajasoluun (Suomen Bioteollisuus, 2003).

Rekombinaatio on tapahtuma, jossa perintötekijöiltään erilaiset solut tai yksilöt tuottavat vanhempiin nähden uudentyyppisen jälkeläisen (Heikkinen *& al.*, 2002).

Restriktioentsyymi on bakteriofageja bakteereissa pilkkova entsyymi. Se tunnistaa tietyn, yleensä palindromisen, DNA-sekvenssikohdan ja katkaisee kaksisäikeisen DNA-sekvenssin kyseisestä kohdasta. Katkaistu kaksisäikeinen DNA voi olla joko tylppä- tai tahmeapäinen (Berg *& al.*, 2002).

Sovitus (alignment, kohdistus, linjaus, rinnastus) tarkoittaa sekvenssien vastinmerkkien tai vastinkohtien asettamista kohdakkain (Heikkinen *& al.*, 2002).

Sekvensointi (sequencing) tarkoittaa nukleotidi- tai aminohappoketjun sekvenssin kokeellista määrittämistä. DNA:n sekvensoinnilla tarkoitetaan DNA:n emäsjärjestyksen määrittämistä (Heikkinen *& al.*, 2002).

Tahmeapäinen (sticky ended) tarkoittaa kaksisäikeisen DNA:n päätä, jossa on katkaisukohdan epäsymmetrisyydestä johtuen lyhyt yksisäikeinen kohta. Sen ansiosta säie voi liittyä helpommin toiseen DNA-molekyyliin (Heikkinen *& al.*, 2002).

Templaatti (malli) on tunnettu rakenne, joka ohjaa toisen rakenteen syntyä. Se voi olla esim. nukleotidisekvenssi, joka ohjaa vastinnukleotideistä muodostuvan sekvenssin synteesiä (Heikkinen *& al.*, 2002).

Templaatti-kartta -strategia on yksinkertainen, mutta tehokas menetelmä DNA-sanojen suunnitteluun (Arita, 2004).

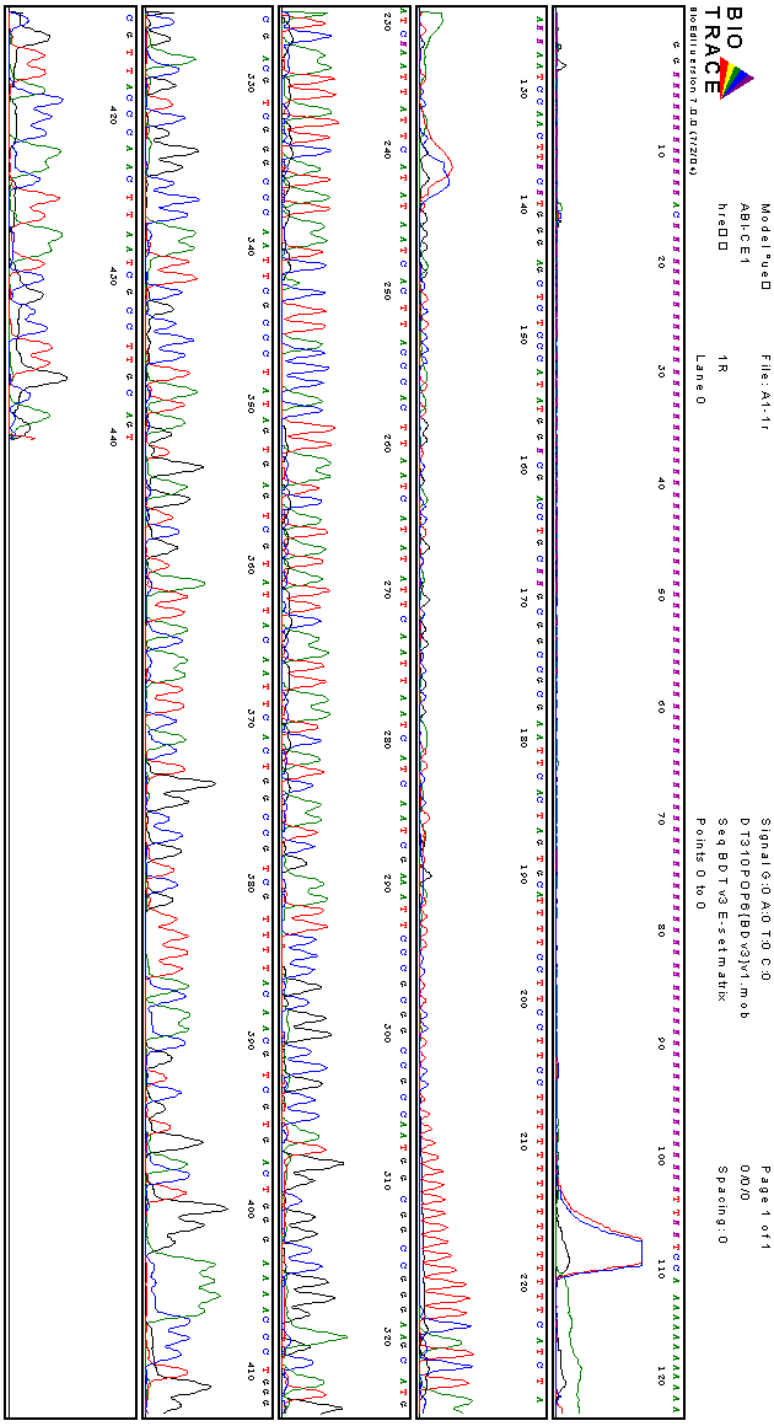
Translaatio (proteiinisynteesi) on DNA:n RNA:ksi kopioimisen jälkeinen tapahtuma, jossa lähettiRNA:n emäsjärjestys käännetään proteiinin aminohappojärjestykseksi (Heikkinen *& al.*, 2002).

Tylppäpäinen (blunt ended) on kaksisäikeisen DNA:n pää, jonka säikeet ovat yhtä pit-

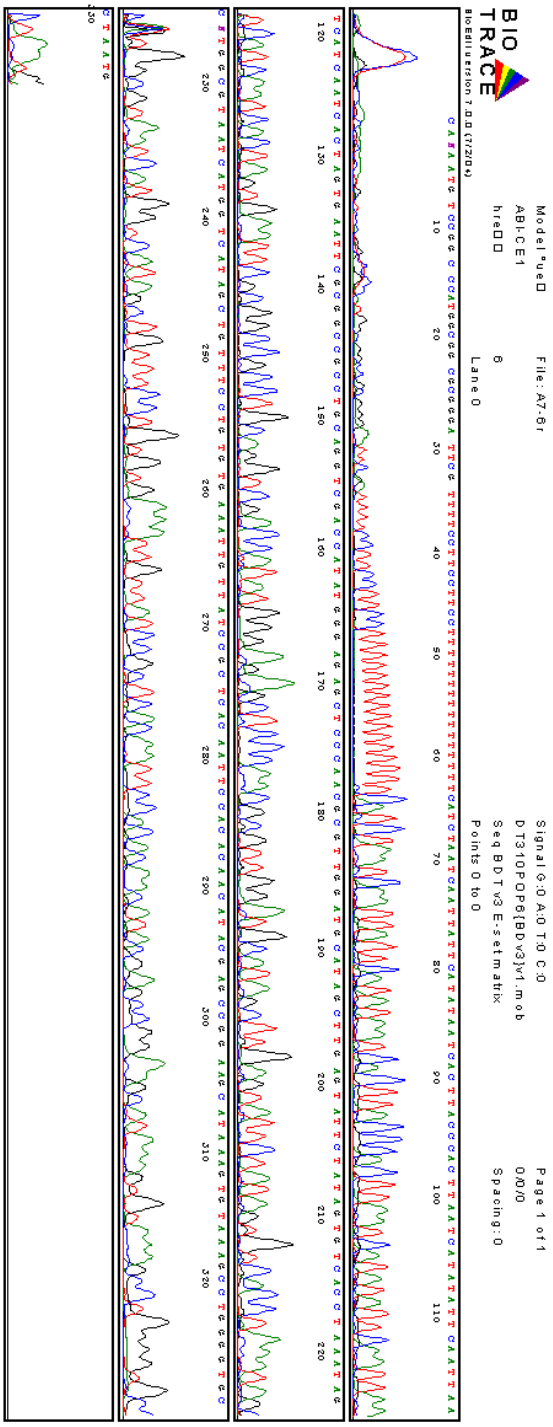
kät (Heikkinen *&* *al.*, 2002).

Vektori on DNA-molekyyli tai bakteerin plasmidi, johon siirrettävä geeni liitetään ja joka kuljettaa geenin vastaanottajasoluun (Suomen Bioteollisuus, 2003).

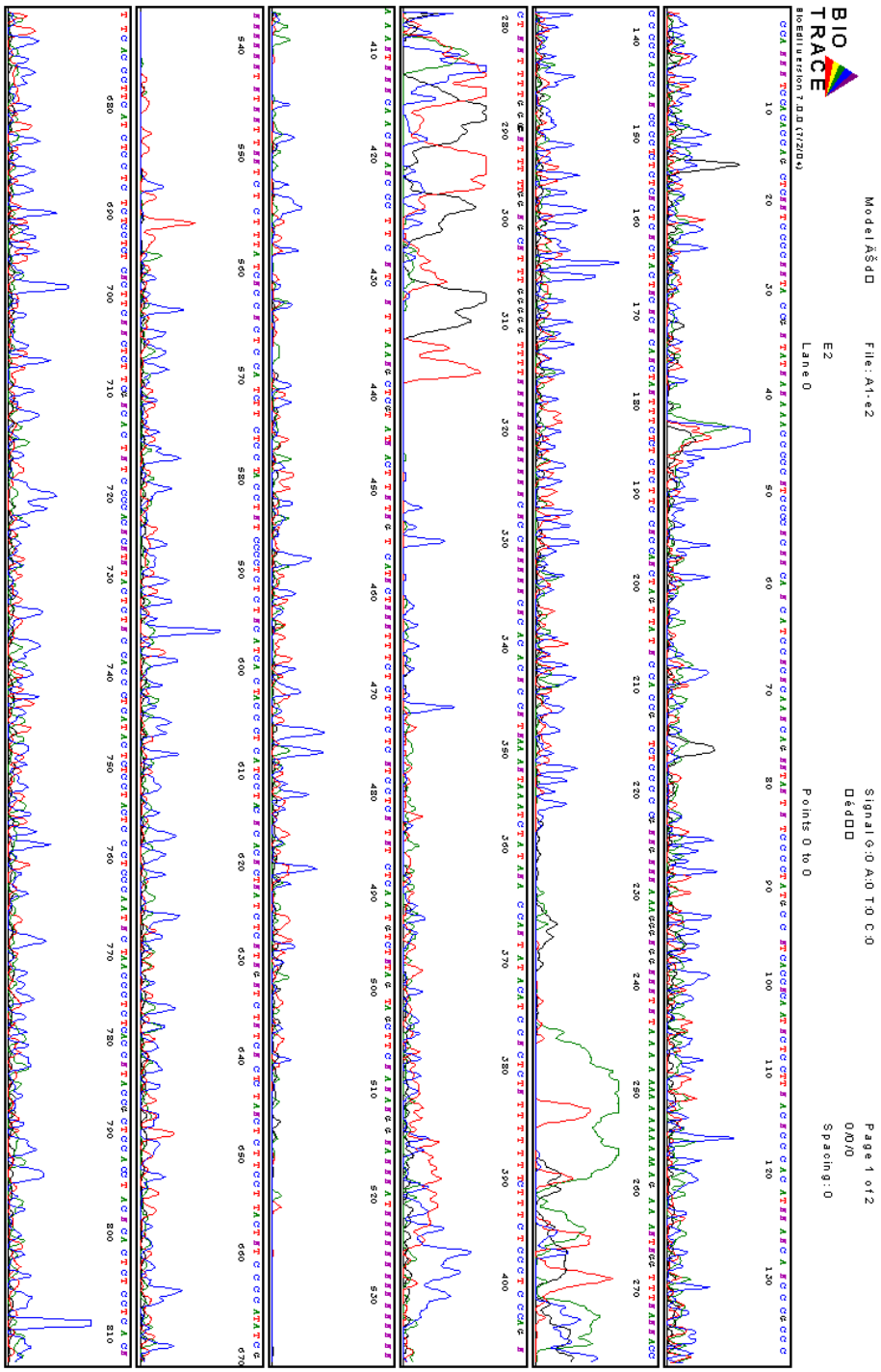
Liite 2: Sekvensoinnin tulokset



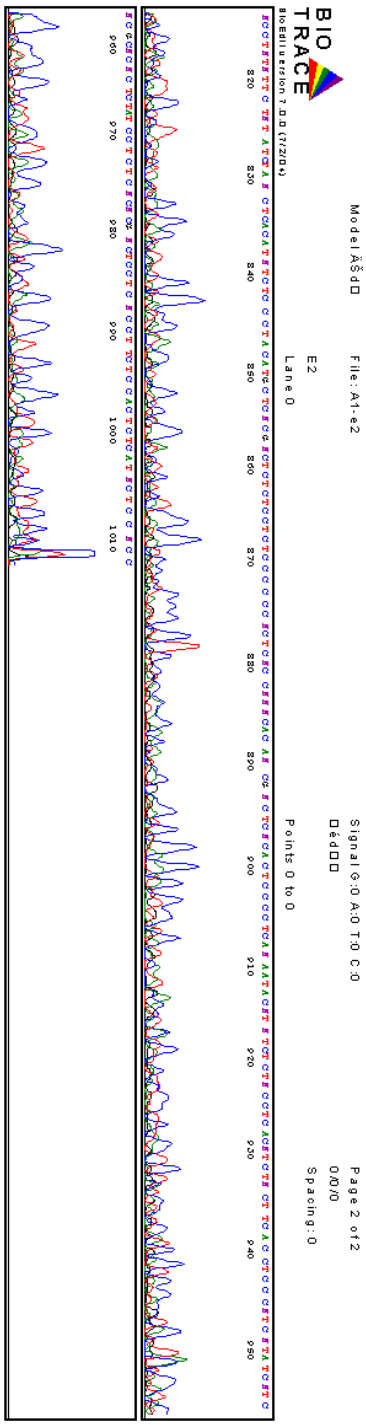
Kuva 1: Pesäkkeestä 1. kerätystä *E. coli* -bakteerikasvustosta sekvensoitu plasmidin DNA.



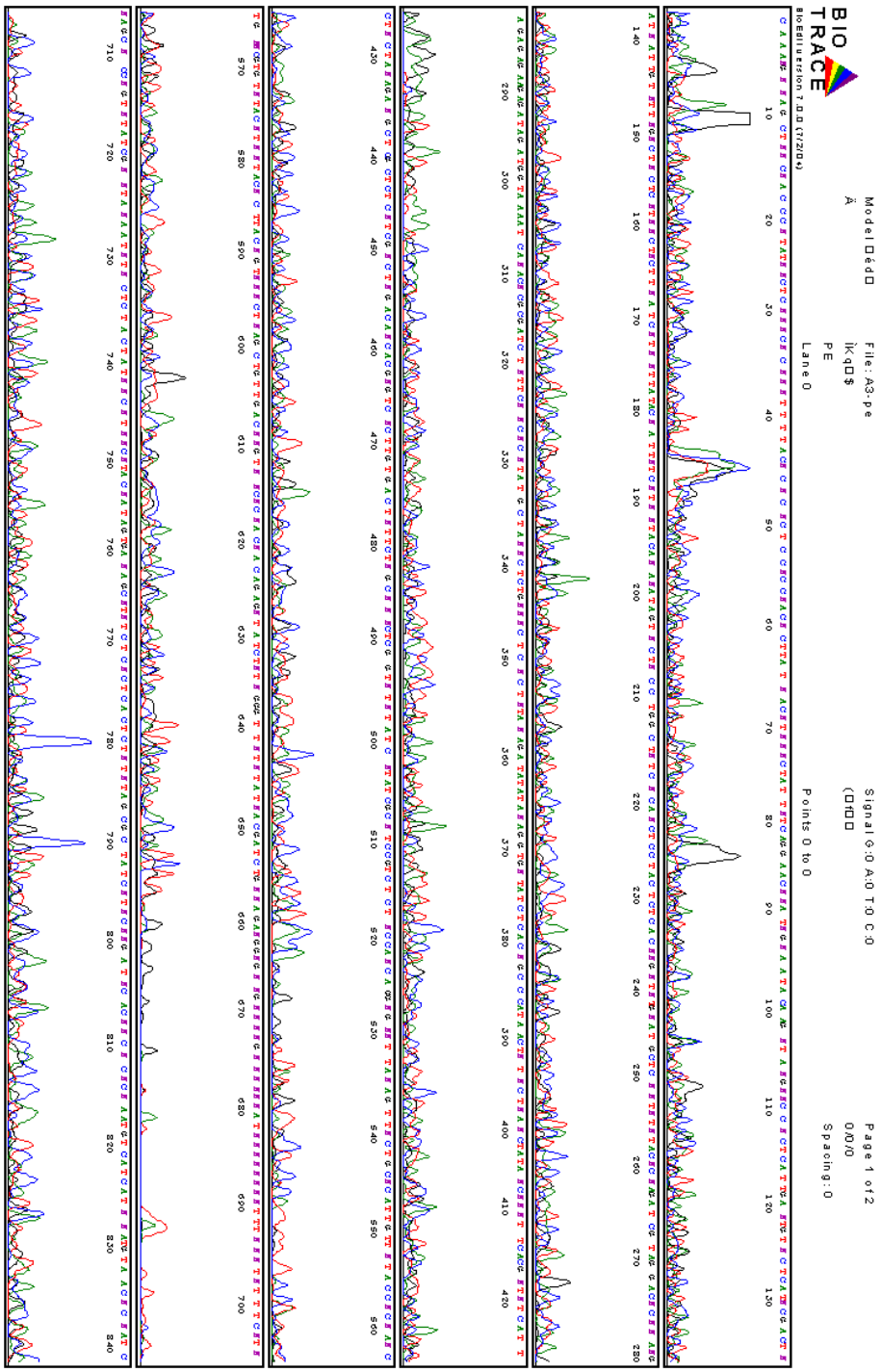
Kuva 2: Pesäkkeestä 6. kerätystä *E. coli* -bakteerikasvustosta sekvensoitu plasmidin DNA.



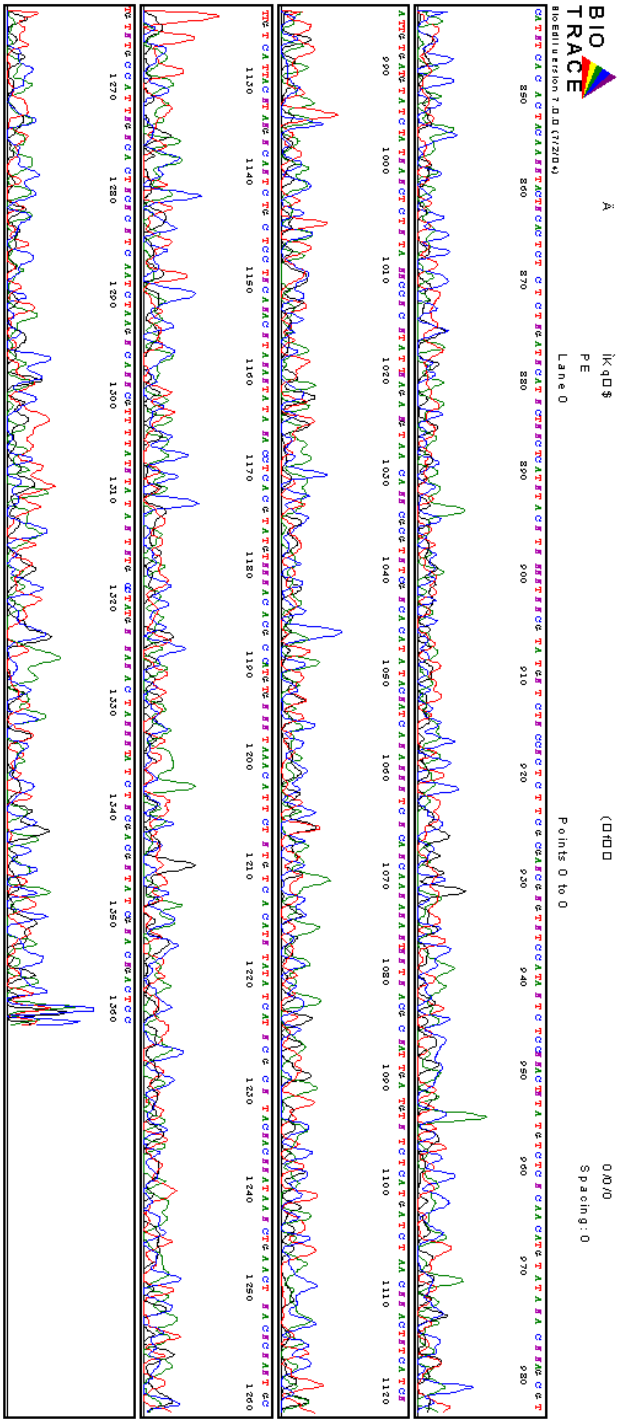
Kuva 3: Keltaisen värin mukaan kalastetun DNA:n sekvenssi (Osa 1).



Kuva 4: Keltaisen värin mukaan kalastetun DNA:n sekvenssi (Osa 2).

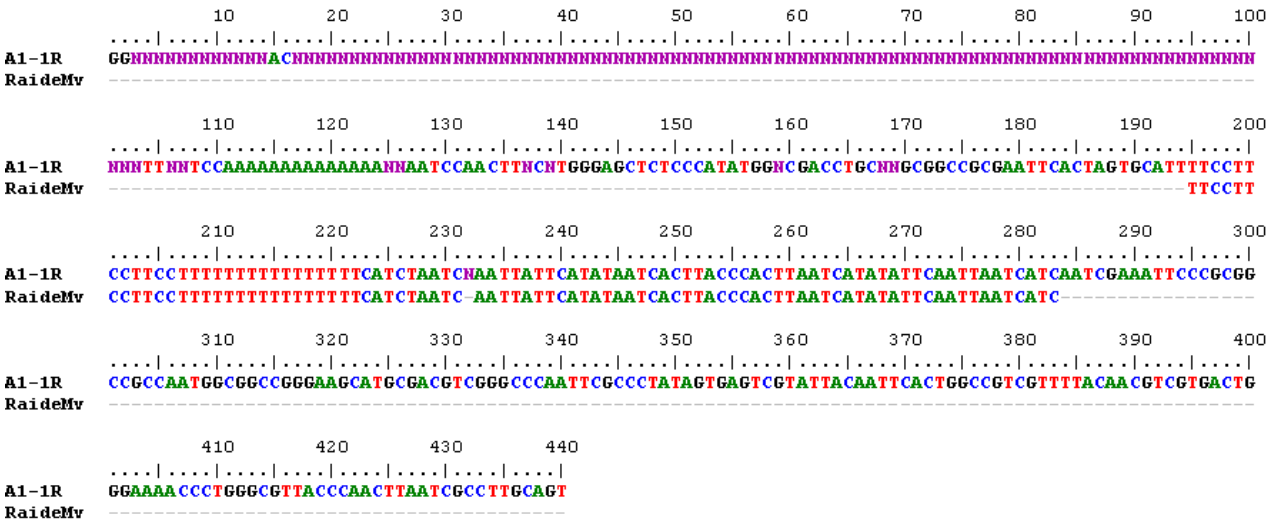


Kuva 5: Punaisen värin mukaan kalastetun DNA:n sekvenssi (Osa 1).

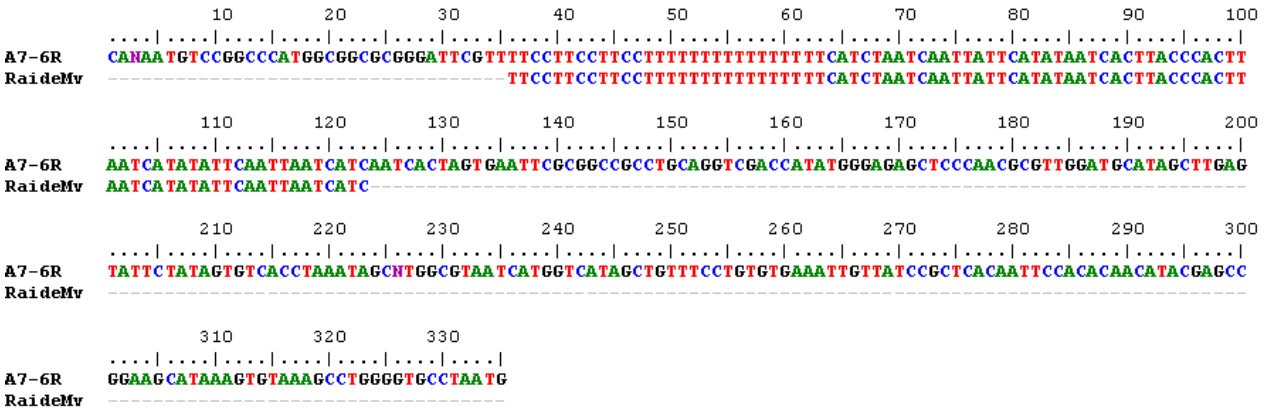


Kuva 6: Punaisen värin mukaan kalastetun DNA:n sekvenssi (Osa 2).

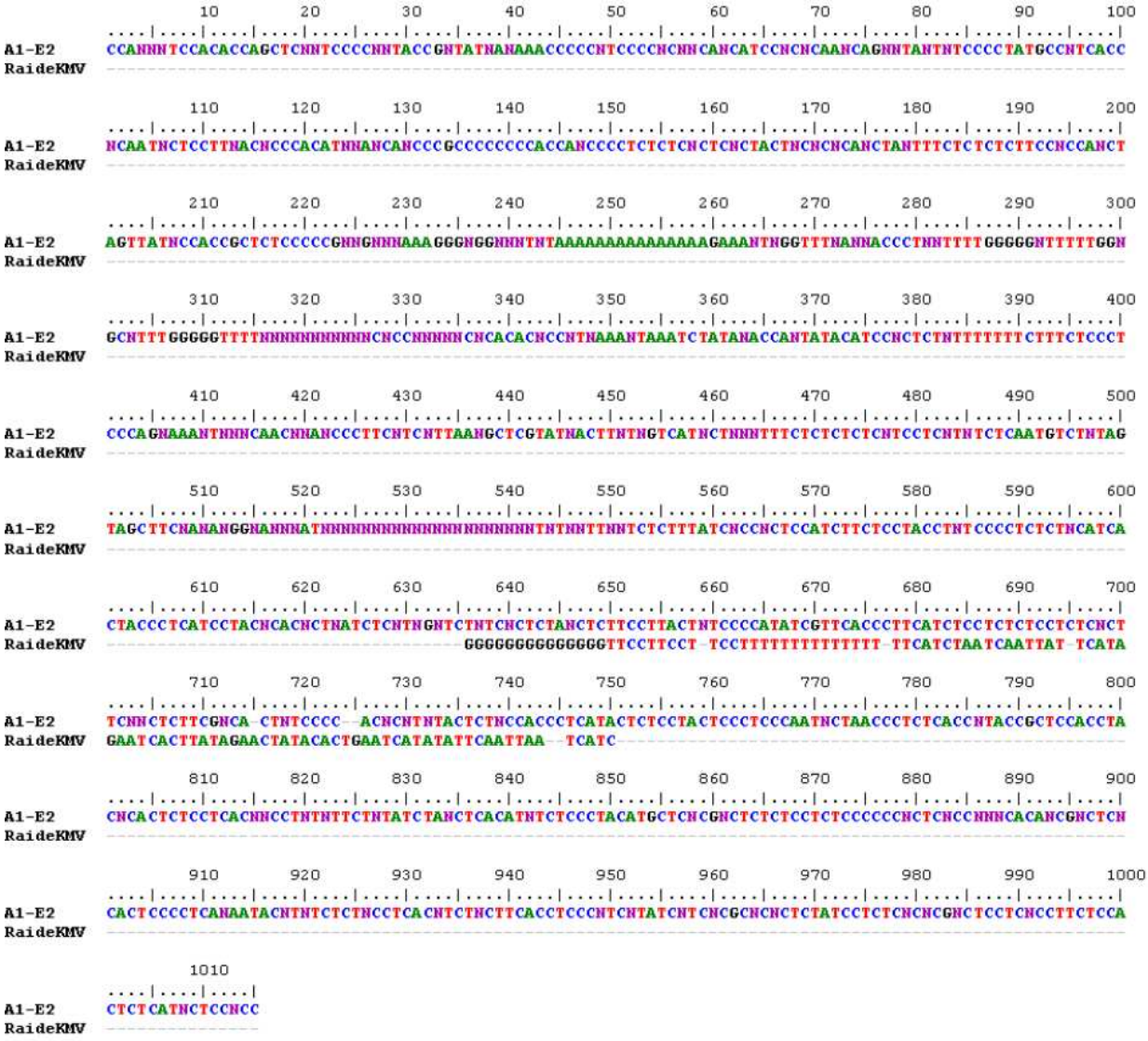
Liite 3: Sekvenssien sovitukset



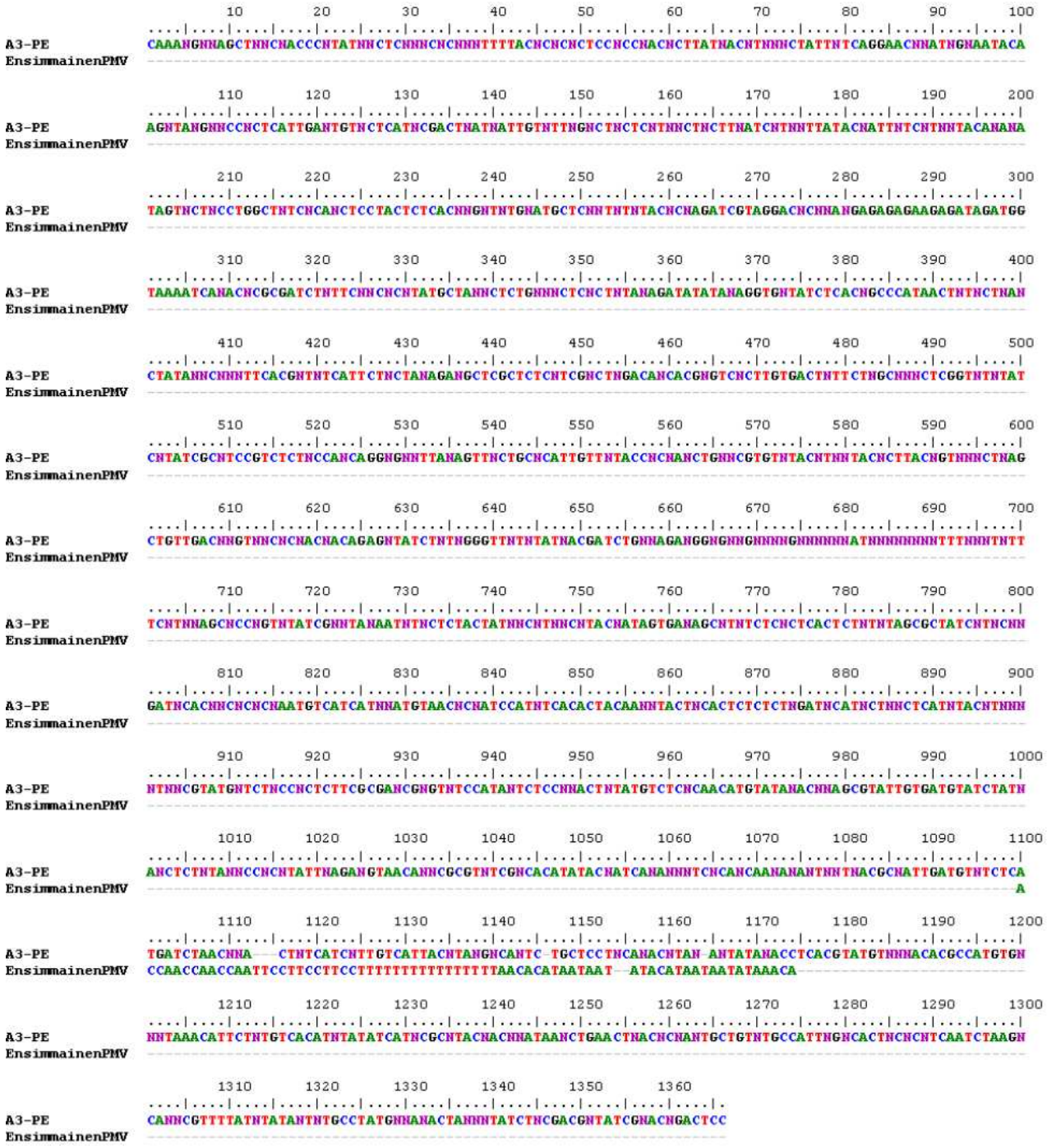
Kuva 1: A1-1R:n ja RaideMV-kuvan sekvenssien sovitus.



Kuva 2: A7-6R:n ja RaideMV-kuvan sekvenssien sovitus.



Kuva 3: A1-E2:n ja RaideKMV-kuvan sekvenssien sovitus.



Kuva 4: A3-PE:n ja EnsimmainenPMV-kuvan sekvenssien sovitus.