

Timo Stadius

**MAHDOLLISUUDET JA HAASTEET DATAN
TALLENTAMISESSA DNA-KETJUIHIN**



JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN
TIEDEKUNTA

2018

TIIVISTELMÄ

Stadius, Timo

Mahdollisuudet ja haasteet datan tallentamisessa DNA-ketjuihin

Jyväskylä: Jyväskylän yliopisto, 2018, 35 s.

Tietojärjestelmätiede, Kandidaatin tutkielma

Ohjaaja(t): Taipalus, Toni

Tämän tutkimuksen tarkoituksena on luoda katsaus mahdollisuuksista ja haasteista, joita datan tallentamiseen DNA-ketjuihin liittyy sekä luoda uutta tietoa tiedeyhteisölle. DNA-ketjujen käyttäminen ulkoisen datan tallentamiseen voi mahdollistaa massiivisia datavarastoja hyvin kompaktiin ja pitkän säilymisajan omaavaan tilaan. Suuren kapasiteetin ja pitkän säilymisajan omaavalle tallennusmenetelmälle on syntynyt viime aikoina tarve kasvavien datamäärien vuoksi. DNA-ketjuihin tallentamalla tämä on mahdollista saavuttaa virheettömästi, johon muiden tallennusmenetelmien avulla ei kyetä. Vastapainona DNA-ketjuihin tallentamisen kustannukset ovat tällä hetkellä varsin korkeat eikä tallennettu data ole luettavissa kovinkaan nopeasti. Jotta DNA-ketjujen soveltuvuudesta ulkoisen datan tallennusmenetelmänä voitaisiin saada parempi käsitys, on sitä hyvä verrata johonkin jo olemassa olevaan, kuten nykyisiin tiedontallennusmenetelmiin. Näin ollen saadaan tarkempi kuvaus siitä, onko DNA-ketjuihin tallentaminen mahdollinen ratkaisu datan varastointiin niihin teknologioihin verrattuna, joita nykyään käytämme. Tämä tutkielma on toteutettu kirjallisuuskatsauksena käyttäen hyväksi mahdollisimman laadukasta sekä relevanttia lähdekirjallisuutta.

Asiasanat: DNA-ketju, tiedon hallinta, kiintolevy, kapasiteetti, säilyvyys

ABSTRACT

Stadius, Timo

Possibilities and challenges of storing data in DNA-chains

Jyväskylä: University of Jyväskylä, 2018, 35 p.

Information Systems Science, Bachelor's Thesis

Supervisor(s): Taipalus, Toni

The purpose of this thesis is to create a look to the possibilities and challenges regarding storing data in DNA-chains as well as contributing the scientific community. Storing foreign data in DNA-chains could enable massive data storages in a small-sized and long-lasting way. In recent years there's been demand for such a storage of data since the amount of data produced nowadays is reasonably vast. By storing data in DNA-chains it is also possible to make it error-free, which no other storing method available is unable to provide. For the contrary, DNA-chains do lack in cost-efficiency and reading speed, when compared to the methods being used nowadays. To get a deeper and better picture on if DNA-chains do contribute as a foreign data storage, it is relevant to compare it to the methods currently existing. By that it is easier to see if it is a solution compared to existing methods of storing data. This thesis is conducted as a literature review, which aims to use high-class source material relevant to the topic of this thesis.

Keywords: DNA-chain, data storage, disk drive, capacity, longevity

KUVIOT

KUVIO 1 DNA-ketjun rakenne	13
KUVIO 2 Osittainen kemiallinen rakenne	14

TAULUKOT

TAULUKKO 1 Tallennusmenetelmien kapasiteetti ja tallennustiheys	18
TAULUKKO 2 Tallennusmenetelmien kustannukset.....	25
TAULUKKO 3 Tallennusmenetelmien lukunopeudet	27

SISÄLLYS

TIIVISTELMÄ	2
ABSTRACT	3
KUVIOT	4
TAULUKOT	4
SISÄLLYS.....	5
1 JOHDANTO.....	6
2 NYKYISET TIEDONTALLENNUSMENETELMÄT.....	8
2.1 Tiedontallennusmenetelmien historiaa	8
2.2 Nykyiset tiedontallennusmenetelmät.....	11
3 DNA-KETJU JA SEN KÄYTTÖ DATAN TALLENTAMISEEN	12
3.1 DNA-ketjun määritelmä	12
3.2 Datan tallennus DNA-ketjuun.....	15
4 MAHDOLLISUUDET DATAN TALLENTAMISESSA DNA-KETJUIHIN NYKYISIIN TIEDONTALLENNUSMENETELMIIN NÄHDEN.....	17
4.1 Kapasiteetti	17
4.2 Säilyvyys	20
4.3 Virheettömyys	22
5 ONGELMAT DATAN TALLENTAMISESSA DNA-KETJUIHIN NYKYISIIN TIEDONTALLENNUSMENETELMIIN NÄHDEN.....	25
5.1 Kustannukset.....	25
5.2 Lukunopeus.....	27
6 YHTEENVETO	29
LÄHTEET	31

1 Johdanto

Vuonna 2007 ihmiskunta varastoi 290 eksatavua (2320 miljoonaa terabittiä yhden tavun vastatessa 8 bittiä) dataa ja vuosittain tämä määrä kasvaa 23 %:lla (Hilbert & López, 2011). Tällaisten tietomäärien varastointiin pyritään löytämään uusia ratkaisuja nykyisten tiedontallennusmenetelmien sijasta. Yhtenä vaihtoehtona on esitetty DNA-ketjuja niissä piilevien mahdollisuuksien vuoksi. Tässä tutkielmassa tullaan käymään läpi DNA-ketjujen mahdollisuuksia sekä haasteita vaihtoehtoisena datan tallennusmenetelmänä.

DNA-ketju on jo luonnostaan menetelmä varastoida dataa, sillä jokaisen elollisen olennon DNA sisältää kaiken kyseiseen olentoon liittyvän informaation, pakaten sen samalla hyvin pieneen ja tiiviiseen tilaan. Datan tallentamiseksi on ehdotettu useita erilaisia ratkaisuja. Toiset ovat esittäneet kahden bitin tallentamista per nukleotidi, toiset yhden bitin tallentamista (Church, Gao & Kosuri, 2012; Blawat, Gaedke, Hütter, Chen, Turczyk, Inverso, Pruitt & Church, 2016; Erlich & Zielinski, 2017). Viimeisimmät tutkimukset ovat kuitenkin osoittaneet, että yhden nukleotidin maksimaalinen kapasiteetti on alle kaksi bittiä (Erlich & Zielinski, 2017).

Tutkielman tarkoituksena on selvittää, millä tavoin DNA-ketjut voisivat tarjota vastauksen suurten datamäärien varastointiin liittyviin ongelmiin. Ongelmaa tullaan tarkastelemaan kahden tutkimuskysymyksen kautta;

- Millaisia mahdollisuuksia DNA-ketjuihin tallentamisella on verrattuna nykyisiin tiedontallennusmenetelmiin?
- Millaisia haasteita DNA-ketjuihin tallentamisella on nykyisiin tiedontallennusmenetelmiin nähden?

DNA-ketjun mahdollisuudet tiedontallennusmenetelmänä perustuvat muun muassa sen kapasiteettihyötyihin, datan säilyvyyteen sekä datan virheettömyyteen liittyviin hyötyihin. Samalla tämä teknologia sisältää vielä haasteita, kuten lukunopeuteen liittyvät haasteet sekä kustannuksiin liittyvät haasteet.

Tutkielma etenee seuraavalla tavalla. Luvussa kaksi tullaan tarkastelemaan digitaalisten tiedontallennusmenetelmien historiaa ja tehdään katsaus

nykyään käytössä oleviin tiedontallennusmenetelmiin. Tarkoituksena ei kuitenkaan ole eritellä näitä metodeja kovinkaan syvällisellä teknologisella tasolla, vaan ennemminkin tehdä yleiskatsaus näiden suhteen. Luvussa kolme tarkastellaan ensin DNA-ketjun rakennetta, jonka jälkeen paneudutaan siihen, kuinka ulkoista dataa voidaan varastoa DNA-ketjuun. Tämän kohdalla käydään läpi muutama erilainen menetelmä, joita on kehitetty sekä se, kuinka itse prosessi tapahtuu.

Neljännessä luvussa paneudutaan tutkimaan ensimmäistä tutkimuskysymystä, eli mitä mahdollisuuksia DNA-ketjuihin tallentamisella voidaan saavuttaa nykyisin käytössä oleviin tiedontallennusmenetelmiin nähden. Luvussa tullaan tarkastelemaan näitä mahdollisuuksia kolmesta eri näkökulmasta: kapasiteettiin liittyvien mahdollisuuksien näkökulmasta, datan säilyvyyden näkökulmasta sekä datan virheettömyyden näkökulmasta. Jokaista näistä näkökulmista tullaan vertaamaan luvussa kaksi esiteltyihin nykyään käytössä oleviin tiedontallennusmenetelmiin.

Tämän jälkeen luvussa viisi vastataan toiseen esitettyyn tutkimuskysymykseen, eli millaisia haasteita datan tallentamisessa DNA-ketjuihin esiintyy verrattuna nykyisiin tiedontallennusmenetelmiin. Tätä kysymystä lähestytään kahdesta eri näkökulmasta. Ensin tarkastellaan datan tallentamisen kustannuksiin liittyviä haasteita, jonka jälkeen vuorossa on datan lukemisen nopeuteen liittyvät haasteet. Myös tässä luvussa suoritetaan vertailu nykyisin käytössä oleviin tiedontallennusmenetelmiin. Tutkielman päättää yhteenvetoluku, jossa kerrataan vielä tutkielmassa käsitellyt asiat, vastaukset tutkimuskysymyksiin sekä esitellään tutkielman aiheesta mahdollisia jatkotutkimuskohteita.

Tutkielma toteutetaan kirjallisuuskatsauksena. DNA-ketjujen soveltuvuutta tiedon tallentamisen menetelmänä on tutkittu jonkin verran, mutta kattavaa vertailevaa tutkimusta nykyisiin tiedontallennusmenetelmiin ei tähän olla kirjallisuuskatsaukseen mennessä vielä tehty. Tutkielman tavoitteena on synnyttää tiedeyhteisölle uutta tietoa sekä tuoda vastaus tutkimusongelmaan kahden edellä mainitun tutkimuskysymyksen kautta.

2 Nykyiset tiedontallennusmenetelmät

Tiedon hallinnan merkitys kasvaa jatkuvasti digitalisoituneessa yhteiskunnassa, jossa syntyvän datan määrä kasvaa jatkuvasti huomattavissa määrin. Uusille menetelmille on huomattava tarve, sillä nykyiset tiedontallennusmenetelmät eivät ole välttämättä pitkäkestoinen ratkaisu tulevaisuutta ajatellen. Tässä luvussa tarkastellaan tiedontallennusmenetelmien historiaa rajoittaen tarkastelu kuitenkin koskemaan vain digitaalisia tiedontallennusmenetelmiä. Myöhemmin eritellään muutama yleisimmistä nykyään käytössä olevista tiedontallennusmenetelmistä, joihin luvussa 3 käsiteltävää DNA-ketjuihin tallentamista tullaan vertaamaan luvussa 4.

2.1 Tiedontallennusmenetelmien historiaa

Tietoa on tallennettu niin pitkään kuin nykyihminen on ollut olemassa. Jo ensimmäisiä luolamaalauksia voidaan pitää jo jonkinlaisena yrityksenä säilöä tietoa myöhempiä aikoja varten. Tästä ollaan kehitytty ensimmäisiin kirjoitustekniikoihin, kuten nuolenpääkirjoitus, aina papyruksiin ja kirjoihin saakka. Seuraava askel tästä on ollut tiedon tallentaminen digitaaliseen muotoon. Tässä alaluvussa tarkastellaan nimenomaan digitaalisten tiedontallennusmenetelmien historiaa.

Ensimmäiset tietokoneet syntyivät toisen maailmansodan aikana saksalaisen Konrad Zusen ja englantilaisen Alan Turingin toimesta (Bauer & Wössner, 1972; Copeland, 2004). Ensimmäiset kaupalliset koneet ilmestyivät sodan jälkeen, kun IBM toi markkinoille IBM 603 -laitteen 1946. Paria vuotta myöhemmin markkinoille ilmestyi hieman käytännöllisempi ja hienostuneempi IBM 604. Laitte käytti tiedon tallentamisessa hyväksi reikäkortteja. (Pugh & Heide, 2013). Reikäkortit olivat ensimmäinen teknologia, jolla tietoa saatiin varastoitua nopeammin ja tehokkaammin, kuin käsin kirjoittamalla. Teknologia oli kehitetty jo 1890-luvulla (Gray, 1996). Reikäkorttien toiminta perustuu pahvikortteihin, joihin painetaan ennalta määritettyihin kohtiin reikiä, joita kone osaa tällöin lukea

(Kilgour, 1997). Ongelmina reikäkorttikoneiden kanssa oli muun muassa se, että kortit piti syöttää koneeseen tietyssä järjestyksessä, jotta haluttu tulos saatiin. Lisäksi korttien lukeminen oli hidasta eikä yksi korttiensyöttökierros välttämättä syöttänyt koneelle kaikkea haluttua tietoa. (Senko, Altman, Astrahan & Fehder, 1973).

Magneettinauhat korvasivat reikäkortit pääsääntöisenä tiedontallennusmenetelmänä 1950-luvun alussa. Tällöin IBM esitteli magneettinauhalaitteensa IBM 701 ja IBM 726, jotka pohjustivat tiedontallennusmenetelmien kehittymistä seuraavan vuosikymmenen ajan. (Stevens, 1981). Univac oli puolestaan kehittänyt jo vuoteen 1950 mennessä magneettinauhan, joka kykeni tallettamaan yhtä paljon tietoa kuin 10000 reikäkorttia (Gray, 1996). Muutos reikäkorteista magneettinauhoihin nähdäänkin datan hallinnan ensimmäisenä niin sanottuna sukupolvenvaihdoksena. Magneettinauhojen ongelmina olivat kuitenkin virheidenkorjaus ja tiedon päivittäminen, sillä tietojen päivittämiseksi koko magneettinauha tuli uudelleenkirjoittaa (Stevens, 1981). Lisäksi siitä puuttui ns. random-access - toiminto, joka teki datan etsinnästä hitaampaa muihin saman aikakauden teknologioihin verrattuna, joissa vastaava toiminto oli mahdollinen (Piramanayagam & Chong, 2011). HDD (Hard Disk Drive) -kiintolevyjen ajaessa magneettinauhojen ohi pääsääntöisenä tiedontallennusmenetelmänä, jäivät ne kuitenkin elämään muun muassa varmuuskopiointia varten (Stevens, 1981). Niiden kehitys on myös jatkunut tähän päivään saakka (Katayama, Chinda, Shimizu, Mikami, Suzuki & Noguchi, 2016).

Vuonna 1956 IBM esitteli ensimmäisen magneettisen kiintolevyn eli HDD-kiintolevyn, RAMACin, jonka kapasiteetti oli 5 megabittiä. HDD-kiintolevyjen toiminta perustuu siihen, että pyörii levy, joka toimii magneettisena datavarastona. Sen lisäksi kiintolevyssä on luku- /kirjoituspää, joka vastaa luku- ja kirjoitusoperaatioiden suorittamisesta levyiltä. Lisäksi levyssä on useita piirejä, jotka vastaavat sen toiminnasta, kuten siitä, mitä tietoa levyiltä halutaan lukea tai minne uutta tietoa kirjoitetaan. Kirjoitusoperaatiot voidaan suorittaa levyn molemmille puolille kapasiteetin kasvattamiseksi. Lisäksi kiintolevyssä voidaan käyttää useampaa levyä kapasiteetin lisäämiseksi. Kirjoitusoperaatiot tehdään jokainen omalle ”radalleen” levyllä. Näitä ratoja voi yhdellä puolella levyä olla jopa 42 kilometrin edestä. (Piramanayagam & Chong, 2011). HDD-kiintolevyt olivatkin pitkään ylivoimaisia muihin käytössä oleviin tiedontallennusmenetelmiin verrattuna. (Piramanayagam & Chong, 2011).

Vuonna 1971 markkinoille ilmestyivät disketit, joista tuli yksi merkittävimmistä tiedontallennusmenetelmistä. Ilmestymisensä aikoihin diskettejä pidettiin vallankumouksellisina (Amankwah-Amoah, 2016). Disketit mahdollistivat myös ensimmäistä kertaa ohjelmistojen myymisen. 1970-luvun lopulla esimerkiksi monet tekstinkäsittelyohjelmat olivat tietokonevalmistajien itsensä tekemiä (Amankwah-Amoah, 2016). Disketit mahdollistivat ulkopuolisten ohjelmistojen valmistuksen ja myynnin. 1990-luvun puoliväliin mennessä diskettejä myytiin vuosittain jo yli 5 miljardia kappaletta ympäri maailmaa (IBM, 2015). Yhtenä syynä suuriin myyntimääriin oli se, ettei disketille voinut tallettaa kovin paljoa tietoa, esimerkiksi alkuperäinen disketti oli kapasiteetiltaan vain

80 kilotavua eli 640 kilobittia, jota tuohon aikaan pidettiin massiivisena (Amankwah-Amoah, 2016). Vuoteen 2000 mennessä kapasiteetti oli noussut merkittävästi, mutta esimerkiksi optisilla tiedontallennusmenetelmillä kapasiteetit olivat huomattavasti tätä suurempia (Amankwah-Amoah, 2016). Disketit jäivät lopulta muiden tiedontallennusmenetelmien varjoon ja ne tuomittiin kuolleeksi usean eri tahon toimesta 2010-luvulla (Amankwah-Amoah, 2016; Ulanoff, 2010).

Yksi merkittävimmistä tiedontallennusmenetelmistä, joka syrjäytti disketit, olivat optiset tiedontallennusmenetelmät, joista ensimmäisenä markkinoille ilmestyivät CD-levyt. CD-levyjen kehitystyö alkoi monen yrityksen toimesta jo 1970-luvun alkupuolella (Peek, 2010). Yhtenä suurimmista toimijoista CD-levyjen kehityksen suhteen toimi Phillips, joka tuotti ensimmäisen CD-levyn prototyypin (Peek, 2010). Vuonna 1980 Phillips ja Sony loivat yhdessä standardin CD-levyille ja vuonna 1982 CD-levyt tulivat markkinoille (Peek, 2010). Yksi merkittävistä eduista, mikä CD-levyillä on disketteihin nähden, on niiden kapasiteettihyöty, sillä CD-levylle kyetään tallentamaan yli 700 megatavua eli 5,6 gigabittia dataa verrattuna diskettien reiluun kahdeksaan megabittiin (Hilbert & López, 2011). CD-levyjen jälkeen seuraavana optisena tiedontallennusmenetelmänä pinnalle nousivat DVD-levyt. DVD-levyt ovat kymmenen eri yhtiön muodostaman allianssin yhteinen formaatti, joka julkistettiin vuonna 1995 (Kablaw, 1996). Markkinoille DVD-levyt tulivat vuonna 1996 (Immink, 1996). Etuina CD-levyihin, DVD-levyillä oli moninkertainen tallennuskapasiteetti, kuten myöhemmin tässä tutkielmassa huomataan. Levy oli lisäksi myös huomattavasti ohuempi, kuin CD-levy (Immink, 1996). Lisäksi lukemiseen ja kirjoittamiseen käytetään pienempää laseria, joka mahdollistaa suuremman tallennustiheyden (Michigami & Tanaka, 1998). Viimeisimpänä optisena tiedontallennusmenetelmänä markkinoille on ilmestyneet Blu-ray -levyt vuonna 2003 (Miyagawa, 2014). Blu-ray -levyjen kehitystyö alkoi jo 1990 -luvun alussa tutkijoiden löydettyä mahdollisuuden tallentaa dataa entistä tiheämmin lasereiden avulla (Miyagawa, 2014). Blu-ray -levyt olivat jälleen askel eteenpäin optisten tallennusmenetelmien kentällä nostaten jälleen rimaa tallennuskapasiteetin ja tallennustiheyden suhteen, kuten myöhemmin tässä tutkielmassa käy ilmi. Lisäksi myös virheidensietokyky parani DVD-levyihin verrattuna (Blu-ray Disk Founders, 2010).

Optisten tiedontallennusmenetelmien samoihin aikoihin kehitettiin myös muitakin tiedontallennusmenetelmiä. 1980-luvun lopulla esiteltiin Flash-muisti, joka on niin sanottu puolijohdemuisti (Bez ym., 2003). Flash-muistien kehityksen voidaan kuitenkin katsoa käynnistyneen jo 1960-luvun lopulla (Coughlin, 2017). Tässä tiedon tallentamiseen käytetään erityisiä soluja, joihin kirjoittaminen, lukeminen ja muokkaaminen tapahtuu sähköisesti, joka puolestaan on nopeampi tapa kuin esimerkiksi magneettiset tai optiset tiedontallennusmenetelmät. Markkinoilla tulosta ei syntynyt ennen 2000-lukua, jolloin menetelmä oli todettu luotettavaksi ja toimivaksi (Bez ym., 2003). Flash-muistia hyödyntää esimerkiksi SSD (Solid State Drive) -kiintolevyt, jotka pyrkivät tuomaan uuden ja paremman vaihtoehdon HDD-kiintolevyjen rinnalle. SSD-kiintolevyt ilmes-

tyivät ensimmäisen kerran markkinoille SanDiskin toimesta vuonna 1990-luvun vaihteessa (Coughlin, 2017). Niiden etuina suhteessa HDD-levyihin on muun muassa se, ettei niissä ole liikkuvia osia, jotka voisivat rikkoontua. Lisäksi SSD-levyt ovat toiminnaltaan nopeampia kuin HDD-levyt, jonka huomaamme jälleen myöhemmin tässä tutkielmassa.

2.2 Nykyiset tiedontallennusmenetelmät

Edellä käsiteltiin tiedontallennusmenetelmien historiaa. Osa edellä mainituista metodeista on kuitenkin jäänyt uudempien teknologioiden jalkoihin, ja sitä kautta jäänyt unholaan. Luodaan seuraavaksi katsaus näistä tiedontallennusmenetelmistä vielä käytössä oleviin menetelmiin.

Kiintolevyistä HDD-kiintolevyt sekä SSD-kiintolevyt ovat yleisesti käytössä olevia metodeja. HDD-kiintolevyt ovat olleet 1960-luvulta lähtien pääsääntöinen tiedontallennustapa digitaalisissa tietokoneissa, joten 50 vuoden aikana ne ovat luoneet itselleen hyvin vahvan jalansijan markkinoilla (Piramanayagam & Chong, 2011). Flash-muistille on kehitetty monia käyttötarkoituksia, ja se toimiikin pohjana SSD-kiintolevyjen toiminnalle. Flash-pohjaisten muistilaitteiden myynti onkin ollut jatkuvassa nousussa tämän vuosikymmenen aikana 2008 talouskriisiä lukuun ottamatta (Coughlin, 2017). Myös magneettinauhut ovat edelleen käytössä arkistointi- sekä varmuuskopiointitarkoituksia varten (Katayama ym., 2016). Optisista tiedontallennusmenetelmistä käytössä ovat CD-levyt, DVD-levyt sekä Blu-ray - levyt. Näistä CD-levyt ovat menettäneet huomattavasti markkinoita, ja Yhdysvalloissa ensimmäiset suuret kauppatjetut ovatkin päättäneet lopettaa CD-levyjen myynnin (Christman, 2018).

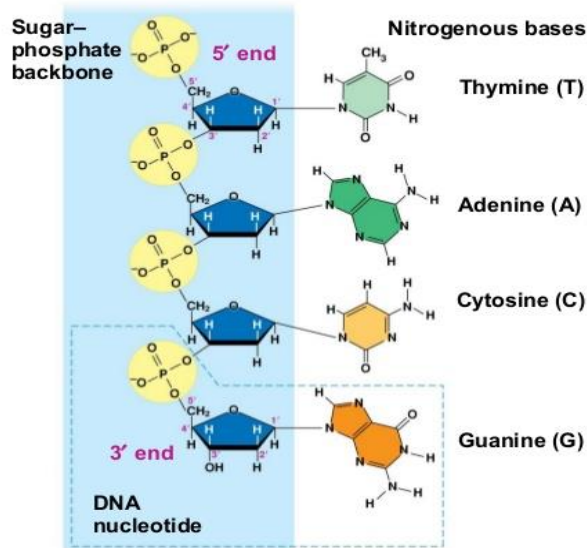
3 DNA-ketju ja sen käyttö datan tallentamiseen

DNA (Deoxiribose Nucleic Acid) on kaiken elämän rakennusaine. Se sisältää kaiken informaation tämän tutkielman kirjoittajasta, lukijasta, lähdeluettelossa mainitusta tutkijoista tai jopa siitä puusta, josta valmistetulle paperille tämä tutkielma on kirjoitettu. DNA-ketju on siis valtaisa tietovarasto, joka nykyteknologialla voidaan kätevästi ottaa käyttöön myös ulkopuolisen datan varastointia ajatellen. Tässä luvussa määritellään, mitä DNA:lla tarkoitetaan, kuinka se rakentuu sekä miksi DNA on otollinen alusta datan tallentamiselle. Luvussa tullaan tarkastelemaan myös DNA-ketjuihin tallentamista, sen menetelmiä ja toimintaperiaatetta.

3.1 DNA-ketjun määritelmä

Tämä alaluku perustuu suurimmilta osin Campbell, Reece, Urry, Cain, Wasserman, Minorsky & Jackson (2015, s. 54, 371-376, 392-398) teokseen. DNA-ketjun rakenne pysyi pitkään tuntemattomana. Vasta 1950-luvun alussa kaksi tutkijaa ratkaisi DNA:n rakenteen erään toisen tutkimuksen sivussa (Watson & Crick, 1953). DNA-ketjuilla tarkoitetaan kahdesta erillisestä polynukleotidiketjusta rakentunutta, kaksoiskierteelle kiertynyttä rihmastoja, jossa rihmastot kulkevat toisiinsa nähden vastakkaisiin suuntiin. Rihmaston rakenne koostuu sen rangasta sekä siihen kiinnittyneistä nukleinihapoista, nukleotideista, jotka kiinnittyvät toisiaan vasten vetysidoksin, sekä yhden ketjun sisällä toisiinsa fosfaattiryhmien avulla. Yksittäinen nukleotidi rakentuu yhdestä fosfaattiryhmästä, pentoosisokerista nimeltä deoxiriboosi, ja siihen kiinnittyneestä typpiä sisältävästä. Rakennetta voidaan havainnollistaa paremmin seuraavan kuvan avulla. Alla olevassa kuvassa fosfaattiryhmä on kuvattu keltaisella, deoxiriboosi sinisellä typpiä sisältävien ollessa erikseen mainittuina (Kuvio 1).

Figure 13.5



© 2014 Pearson Education, Inc.

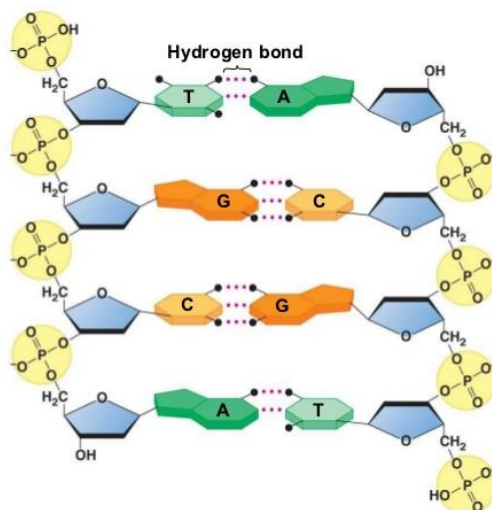
KUVIO 1 DNA-ketjun rakenne

Ketjun rakentumisen täysin ymmärtääkseen on otettava vielä askel syvemmälle nukleotidin rakenteeseen. Deoxiriboosi on rakentunut niin, että se alkaa viisihiilisestä yhdistelmästä, tarkoittaen sitä, että kyseisessä yhdisteessä on viisi hiiliatomia, päättyen kolmihiiliseen yhdistelmään. Tämä on merkittävä tekijä DNA-ketjun rakentumisessa. Nukleotidin fosfaattiryhmä kiinnittyy deoxiriboosin viisihiiliseen yhdistelmään typpiämäksen kiinnittyessä puolestaan tätä seuraavaan hiileen. Ketjun sisällä tätä seuraavan nukleotidin fosfaattiryhmä puolestaan kiinnittyy edeltävän nukleotidin deoxiriboosin kolmihiiliseen yhdistelmään. Näin syntyy DNA-ketjun rankana toimiva sokerifosfaattiranka. Tämän pohjalta voidaan myös päätellä ketjun lukusuunta. DNA-ketju lähtee liikkeelle viisihiilisestä päästä päättyen kolmihiiliseen päähän. Ketjun kanssa vastakkain tuleva ketju asettuu toiseen nähden erisuuntaisesti niin, että toisen ketjun viisihiilistä päätä vasten on toisen ketjun kolmihiilinen pää. Yllä oleva kuva havainnollistaa myös fosfaattiryhmien liitokset 5- ja 3- hiilisiin yhdisteisiin sekä ketjun suunnan.

Ketjun niin sanotun sisäosan rakentavat neljä eri typpiämästä: adeniini, tyymiini, sytosiini sekä guaniini. Yleisesti kyseisistä typpiämäksistä käytetään englanninkielisiä kirjainlyhenteitä A, T, C ja G, joita myös tässä tutkielmassa tullaan käyttämään. Typpiämäkset voidaan yksinkertaistamisen vuoksi käsittää nukleotideiksi, sillä niiden ominaisuudet ovat nukleotidien kesken ainoa erittelevä tekijä. Nämä nukleotidit asettuvat keskenään vastinpareiksi A:T ja G:C. Näistä A sekä G voidaan luokitella pyrimidiineiksi ja T sekä C puriineiksi. Nukleotidit eivät ole kykeneväisiä pariutumaan muuten, kuin edellä mainitulla tavalla johtuen niiden koosta: Puriinit ovat kaksi kertaa leveämpiä kuin pyrimidiinit. Näin olleen puriini - puriini - pari on liian leveä mahtuakseen DNA-ketjuun pyrimidiini - pyrimidiini - parin ollessa puolestaan liian kapea ketjun vaatimuksiin nähden. Vastinparit yhdistyvät toisiinsa vetysidoksin niin, että jokaisen A:T - parin kaksi vetyatomia muodostavat keskenään vetysidoksen, ja

jokaisen G:C - parin kolme vetyatomia muodostaessa keskenään vetysidoksen. Alla oleva kuva pyrkii havainnollistamaan tätä paremmin. Fosfaattiryhmät ovat jälleen kuvattu keltaisella deoxiriboosin ollessa jälleen sininen. Kuvasta voidaan huomata myös puriinien ja pyrimidiinien välinen kokoero, joka estää näiden pariutumisen keskenään. Typpiemästen väliset vetysidokset on kuvattu piste-
viivoin (Kuvio 2).

LE 10-03db



Partial chemical structure

Copyright © 2005 Pearson Education, Inc. Publishing as Pearson Benjamin Cummings. All rights reserved.

KUVIO 2 Osittainen kemiallinen rakenne

Yhtä kahden nukleotidin välistä yhteenliittymää kutsutaan nukleotidipariksi. Yksittäinen DNA-ketju voi sisältää miljoonia tai miljardeja nukleotidipareja. Jokaisessa yksittäisessä ihmisen solussa on noin 6 miljardia nukleotidiparia.

Näiden nukleotidien pohjalta rakentuu DNA-ketjun koodi. Seuraavaksi perehdytään siihen, kuinka DNA-ketju käytännössä toimii datavarastona. Tärkeässä roolissa tämän suhteen on RNA-ketju, joka rakentuu DNA-ketjun pohjalta. Erona RNA:n ja DNA:n välillä on se, että RNA on rakentunut riboosista deoxiriboosin sijaan sekä se, että typpiemäksistä RNA käyttää tymiinin sijasta urasiilia. RNA ei myöskään rakennu kahdesta vastakkain asettuvasta ketjusta, vaan on yksittäinen nukleotideista rakentunut ketju. RNA-ketjun rakentumista DNA-ketjusta käytetään termiä transkriptio. Tässä DNA-ketjuun kiinnittyy molekyyli nimeltä RNA-polymeraasi. Tämä molekyyli osaa lukea DNA-ketjusta tietyn sekvenssin, tietyn osan yhdestä ketjusta, johon se kiinnittyy. Kiinnittyään RNA-polymeraasi avaa DNA-ketjun liikkuen DNA-ketjua eteenpäin 5-hiilisestä päästä 3-hiiliseen päähän, jossa tietyssä kohtaa DNA-ketjua tarvittava sekvenssi päättyy, ja RNA-polymeraasi irtoaa DNA-ketjusta. Liikkuessaan ketjua eteenpäin, RNA-polymeraasi kirjoittaa DNA-ketjun perusteella RNA-ketjua samalla sulkien DNA-ketjun uudelleen perässään.

RNA-ketjun ollessa valmis, muokataan siitä erikseen mRNA-ketjua, eli message-RNA-ketjua. Tässä alkuperäiseen RNA-ketjuun liitetään alkuun yksi kappale G-nukleotideja sekä loppuun 50-250 kappaletta A-nukleotideja. Näiden

tehtävänä on ilmoittaa, mitkä osat mRNA-ketjusta tullaan käyttämään edelleen. Tämän jälkeen niin sanotusti käyttökelpoiset osat mRNA-ketjusta jaetaan kolmen nukleotidin sarjoiksi, joita kutsutaan kodoneiksi. Näiden kodonien pohjalta muodostuvat puolestaan tietyt aminohapot, joista edelleen rakentuvat proteiinit, jotka puolestaan ovat suuressa vastuussa eliön rakenteesta.

3.2 Datan tallennus DNA-ketjuun

Perehdytään seuraavaksi siihen, kuinka DNA-ketju voisi toimia varastona myös sen kannalta ulkopuoliselle datalle. Edellisessä alaluvussa todettiin, että DNA-ketju rakentuu nukleotideista. Nämä nukleotidit puolestaan ovat vastuussa RNA-ketjun rakentumisesta, joka puolestaan vastaa proteiinien valmistuksesta ja lopulta eliön toiminnasta. Nämä samaiset nukleotidit ovat myös hyvin merkittävässä roolissa, kun käsitellään datan tallentamista DNA-ketjuihin.

Teoriassa tämä toimii niin, että haluttu data avataan binäärimuodossa, josta se muunnetaan vastaamaan DNA-ketjun nukleotideja. Tähän on kehitetty metodi, jossa binääriluku 0 vastaisi nukleotideista joko kirjainta A tai C, ja binääriluku 1 puolestaan vastaisi nukleotideista kirjainta G tai T (Church ym., 2012). Samaa tekniikkaa ehdotetaan myös toisessa tutkimuksessa (Blawat ym., 2016). Tässä tutkimuksessa esitetään myös toinen tapa muuttaa binäärinen data vastaamaan nukleotideja. Erona edelliseen metodiin, jossa yksi nukleotidi vastaisi vain joko lukua yksi tai nolla, tässä metodissa tarkoitus olisi mahdollistaa kaksi lukua yhteen nukleotidiin. Tämä tapahtuisi niin, että binäärilukupari 00 vastaisi nukleotidia A, binäärilukupari 01 vastaisi nukleotidia C, binäärilukupari 10 vastaisi nukleotidia G ja binäärilukupari 11 vastaisi täten nukleotidia T (Blawat ym., 2016). Tämä mahdollistaisi tiiviimmän datan pakkaamisen yhden binääriluvun järjestelmään verrattuna. Toisaalta perusteina yhden binääriluvun tallentamiselle per nukleotidi on käytetty sitä, että täten vältetään hankalasti luettavia DNA-sekvenssejä, joissa yhtä tiettyä nukleotidia olisi useampi kappale perätysten, tai G:C - nukleotidien yhdistelmiä (Church ym., 2016). Perättäisten nukleotidien aiheuttamaan ongelmaan on pyritty luomaan sääntö, jonka mukaan sekvensoitavassa ketjussa ei saisi olla perätysten kuin maksimissaan 3 samaa nukleotidia (Blawat ym., 2016). Toisaalta on myös esitetty, ettei kahden bitin tallentaminen yhteen nukleotidiin ole edes mahdollista (Erlich & Zielinski, 2017). Heidän tutkimuksensa mukaan yhden nukleotidin teoreettinen tallennuskapasiteetti on 1,98 bittiä, ja käytännössäkin 1,60 bittiä. Tutkimuksessa oli käytetty hyväksi video - ja suoratoistoon käytettävää algoritmia, jolla mahdollistettiin aikaisempaa tiiviimpi tapa varastoida dataa. Kasvua aikaisemmin käytössä olleisiin algoritmeihin oli jopa 60 % (Erlich & Zielinski, 2017).

Perehdytään seuraavaksi tekniikkaan, jolla datan varastointi DNA-ketjuun mahdollistetaan. Alussa haluttu data tulee muuntaa binääriseen muotoon. Tämä toteutetaan niin, että testataan mahdollisia yhdistelmiä niin, että saadaan karsittua sellaiset vaihtoehdot, jotka voisivat aiheuttaa ongelmia ja virheitä dataa muutettaessa DNA-ketjun muotoon (Blawat ym., 2016). Tällaisia ovat muun

muassa yli kahden peräkkäisen saman nukleotidin sarjat (Goldman ym., 2013; Blawat ym., 2016; Erlich & Zielinski, 2017). Tämän yhteydessä pitkät koodipätkät katkotaan lyhyemmiksi virheiden edelleen karsimiseksi ja niihin liitetään alkuun tunnisteavaimet tiedon etsinnän helpottamiseksi sekä loppuun virheenkorjauskoodit (Blawat ym., 2016). Lisäksi testataan, ettei muunnosvaiheissa datasta DNA-kejuksi ja takaisin pääse syntymään virheitä lisäämällä satunnaiseen paikkaan jotain nukleotidia vastaava kappale X. Tällöin nähdään muuttuvatko lyhennetyt koodipätkät vastaamaan edellä mainittuja sääntöjä (Blawat ym., 2016). Tämän jälkeen muokataan binäärimuotoinen esitys vastaamaan nukleotideja A, C, G ja T jollain ohjelmointikielillä, esimerkiksi Church ym. (2012) käyttivät tähän Python - ohjelmointikieltä, kuten myös Erlich & Zielinski (2017).

Seuraavaksi nukleotideiksi muutettu koodi lähetetään datamuodossa sekvensointikoneelle, joka muodostaa annetun datan perusteella synteettistä DNA-kejuja (Church ym., 2012; Goldman ym., 2013; Blawat ym., 2016; Erlich & Zielinski, 2017). Näin halutusta datasta saadaan muodostettua nukleotideja ja lopulta synteettisiä DNA-kejuja. Erlich & Zielinski (2017) raportoivat tämän kestävän noin kaksi ja puoli minuuttia tavallisella kannettavalla tietokoneella. Datin lukuoperaatio toimii vastaavalla tavalla. DNA-kejut lähetetään PCR (Polymerase Chain Reaction) -laitteelle, jolla kyetään mm. monistamaan DNA-kejuja, josta ne siirretään tallentamisessakin käytetylle sekvensointilaitteelle, joka kykenee lukemaan DNA-kejuja ja muuttamaan sen datamuotoon python - kieliseksi koodiksi. (Erlich & Zielinski, 2017). Kyseinen operaatio vie noin 9 minuuttia (Erlich & Zielinski, 2017). Tämän jälkeen data on jälleen luettavissa tietokoneella. On myös kehitetty menetelmä, jolla pitkiä DNA-kejuja pystytään tallentamaan ja lukemaan ilman välissä tapahtuvaa sekvensointia, jolloin kustannukset on saatu matalammiksi kuin muissa menetelmissä, mutta samalla tallennuskapasiteetti laskee vain murto-osaan sekvensoimalla rakennetuista DNA-kejuista (Portney, Wu, Quezada, Lonardi & Ozkan, 2008).

4 Mahdollisuudet datan tallentamisessa DNA-ketjuihin nykyisiin tiedontallennusmenetelmiin nähden

Kartoitettaessa uutta teknologiaa, on hyvä eritellä sen mahdollisuudet nykyisiin menetelmiin nähden. Aikaisemmissa luvuissa määriteltiin, mitä tallennusmetodeja yleisesti käytetään datan tallentamiseksi. Tunnistettuja metodeja ovat CD-, DVD- ja Blu-ray - levyt, magneettinauhat, Flash-muisti sekä HDD- ja SSD - kiintolevyt. Tämän jälkeen olemme esitelleet DNA-ketjun ja kuinka sitä voidaan käyttää datan tallentamisen välineenä. Tässä luvussa selvittämään DNA-ketjujen mahdollisuuksia toimia datan tallentamisen alustana tarkastelemalla tätä ensimmäisen tutkimuskysymyksen kautta. Näitä mahdollisuuksia tullaan vertaamaan aikaisemmin esitettyihin nykyisiin käytössä oleviin datan tallennusmenetelmiin. Näitä mahdollisuuksia ovat muun muassa kapasiteettihyödyt sekä virheettömyys.

4.1 Kapasiteetti

Otetaan ensiksi huomioon tallennuskapasiteettiin liittyvät hyödyt. Nykyisellään tallennusmetodien kapasiteetti on hyvinkin rajallinen. DNA-ketjuihin tallennettaessa ei kuitenkaan ole rajoituksia vastaavassa määrin kuin nykyisissä menetelmissä. Eräässä tutkimuksessa on vertailtu tuon ajan kaupallisten tallennusmenetelmien kapasiteettia sekä datan pakkaamisen tiiviyyttä, tässä tapauksessa, kuinka monta bittiä kuutiomillimetrille mahtuu (Church ym., 2012). Tutkijat kuitenkin mainitsevat tehneensä oletuksia esimerkiksi paksuuksien osalta mitaustuloksissaan, näin ollen saadut kapasiteetit ovat teoreettisia. Tulokset on myös tiivistetty niiden havainnollistamiseksi taulukkoon 1.

TAULUKKO 1 Tallennusmenetelmien kapasiteetti ja tallennustiheys

Tallennusmenetelmä	Kapasiteetti	Tallennustiheys
CD-Levy	0,0056 Tb	0,413 Mb / mm ³
DVD-Levy	0,0376 Tb	10,1 Mb / mm ³
Blu-Ray - Levy	0,4 Tb	75, 2 Mb /mm ³
Magneettinauha	1776 Tb	559 Mb /mm ³
Flash-muisti	0,5 Tb	5020 Mb /mm ³
HDD-kiintolevy	80 Tb	3010 Mb /mm ³
SSD-kiintolevy	480 Tb	7,3 Mb /mm ³
DNA (Church ym.)	455 000 Tb / g	5 500 000 000 Mb / mm ³
DNA (Goldman ym.)	455 000 Tb / g	9 500 000 000 Mb / mm ³
DNA (Blawat ym.)	455 000 Tb / g	23 000 000 000 Mb /mm ³
DNA (Erlich & Zielinski)	> 455 000 Tb / g	922 000 000 000 Mb / mm ³

Tiheyksiä mitattaessa on käytetty Church ym. (2012) tutkimuksen arvoja tallennustiheyksien suhteen. Lisäksi muutamat tiheyksistä on erikseen laskettu tähän taulukkoon eri muuttujien vuoksi käyttäen hyväksi niistä tehtyjä tutkimuksia. Kapasiteetit on puolestaan haettu suurimmalta osin eri tutkimuksista, jotta saataisiin taulukkoon luotettavat arvot.

Aloitetaan vertailu CD-levyistä, joiden kohdalla tallennuskapasiteetti on noin 730 megatavua eli 5,6 gigabittiä yhden tavun vastatessa kahdeksaa bittiä, ja tallennustiheys 0,413 megabittiä kuutiomillimetrillä (Hilbert & Lopez, 2011; Church ym., 2012). CD-levyistä hieman kehittyneemmällä DVD-levyillä vastaavat lukemat ovat huomattavasti suuremmat: 4,7 gigatavun eli 37,6 gigabitin tallennuskapasiteetti, tallennustiheyden ollessa 10,1 megabittiä kuutiomillimetrillä (Michigami & Tanaka, 1998; Church ym., 2012). Tästä seuraava askel ovat Blu-ray - levyt. Näiden kohdalla tallennuskapasiteetti on 50 gigatavua eli 400 gigabittiä ja tallennustiheys hieman DVD-levyjä korkeampi 75,2 megabittiä kuutiomillimetrille. (Royon, Bourhis, Bellec, Papon, Bousquet, Deshayes, Cardinal & Canioni, 2010; Church ym., 2012).

Seuraavana ratkaisuna tutkimuksessa esitetään magneettinauha. Magneettinauhaa käytetään muun muassa varmuuskopioiden säilytykseen sen halpuuden ja luotettavuuden vuoksi. Vuonna 2015 Fujifilm ja IBM esittelivät magneettinauhansa, jonka tallennuskapasiteetti on jopa 220 teratavua eli 1776 terabittiä. (Katayama ym., 2016). Church ym. (2012) tekemässä tutkimuksessa käytetyn Oraclen magneettinauhan kapasiteetti on 40 terabittiä ja tallennustiheys 559 megabittiä kuutiomillimetrillä. Seuraavaksi otetaan käsittelyyn Flash-muisti. Tutkimuksessa Flash-muistin kapasiteetiksi saatiin 1,02 terabittiä ja tallennustiheydeksi 5,02 gigabittiä kuutiomillimetrille. (Church ym., 2012). Toisessa tutkimuksessa Flash-muistin kapasiteetiksi arvioitiin 0,5 terabittiä (Fontana, Hezler & Decad, 2012).

Tarkastellaan vielä HDD - ja SSD - kiintolevyjä ennen kuin näitä edellä esiteltyjä verrataan DNA-ketjuun. Church ym. (2012) tutkimuksessaan laskivat

HDD-levyn tallennuskapasiteetiksi 48 terabittiä ja tallennustiheydeksi 3,1 gigabittiä kuutiomillimetrille. Toisessa tutkimuksessa on puolestaan otettu esille kuluttajille saatavilla olevat mallit. Tutkimuksen suurimman HDD-kiintolevyn tallennuskapasiteetti vuonna 2016 on ollut 10 teratavua eli 80 terabittiä (Gonzales, Winkelman, Tran, Sanchez, Woods & Hollywood, 2017). Samaisessa tutkimuksessa suurimman SSD-kiintolevyn tallennuskapasiteetiksi kerrottiin jopa 60 teratavua eli 480 terabittiä. Church ym. (2012) tutkimuksessa ei olla käsitelty SSD-kiintolevyä ollenkaan. Eräässä toisessa tutkimuksessa SSD-kiintolevyn tallennustiheydeksi 120 gigabittiä kuutiotuuma, yksi kuutiotuuma vastaa kuutiomillimetreissä 16385,92101 kuutiomillimetriä, jolloin tiheyden arvo saadaan laskutoimituksella $120\text{Gb}/16385,92101\text{ mm}^3$, joka sieventyy arvoon 7,3 megabittiä kuutiomillimetrillä (Fontana ym., 2012). Sama arvo saatiin myös HDD - kiintolevyille.

Seuraavaksi päästään tutkimaan asiaa DNA-ketjun kannalta. On arvioitu, että yhteen grammaan DNA:ta voitaisiin varastoida jopa 455 eksabittiä eli 455000 terabittiä dataa (Church ym., 2012). Sama tulos on saavutettu myös toisessa tutkimuksessa (De Silva & Ganedoga, 2016). Verrattuna suurimman kapasiteetin omaaviin kiintolevyihin, eroa on jopa 45500% HDD-kiintolevyyn ja SSD-kiintolevyynkin nähden yli 7500% DNA-ketjun hyväksi. Tallennustiheydeksi tässä tutkimuksessa laskettiin 5,5 petabittiä kuutiomillimetrille. Tämä vastaa tiheydeltään 1,28 petabittiä per gramma (Erlich & Zielinski, 2017, mukaillen Church ym., 2012). Myöhemmin esitettävät tiheyden tulokset on laskettu per gramma, joten lasketaan muuttuja, jolla pystytään arvioimaan gramman ja kuutiomillimetrin suhdetta. Muuttujaksi saadaan $(5,5\text{ Pb}/\text{mm}^3) / (1,28\text{ Pb}/\text{g}) = 4,296$. Täten jatkossa esitetyt tiheydet per gramma kerrotaan tällä muuttujalla. Teoreettiseksi tiheydeksi tutkijat puolestaan arvioivat 680 petabittiä grammalle (Erlich & Zielinski 2017, mukaillen Church ym., 2012). Tämä tarkoittaisi täten $680\text{ Pb} \times 4,29 = 2,9$ eksabittiä kuutiomillimetriä kohti.

Myöhemmässä tutkimuksessa on saavutettu 2,2 petabitin grammaa kohti - tiheys (Goldman, Bertone, Chen, Dessimoz, LeProust, Sipos & Birney, 2013). Tämä vastaa näin ollen $2,2\text{ Pb} \times 4,29 = 9,5$ petabittiä kuutiomillimetrille. Tämänkin yli ollaan kuitenkin jo päästy myöhemmässä tutkimuksessa, jossa ollaan saavutettu 5 petabitin grammaa kohti - tiheyksiä (Blawat ym., 2016). Tämä tarkoittaa puolestaan $5\text{ Pb} \times 4,29 = 23$ petabittiä per kuutiomillimetri. Toisaalta viimeisimmissä tutkimuksissa nämä arvot on saatu näyttämään lähes mitättömiltä. Erlich & Zielinski (2017) tekemässä tutkimuksessa tutkijat käyttivät koodin pakkaamiseen uutta algoritmia, joka mahdollistaa heidän mukaansa 60% tehokkaamman datan pakkaamisen. Tässä tutkimuksessa saavutettiin 215 petabitin tallennustiheys, teoreettisen maksimin noustessa jopa 200 eksabittiin (Erlich & Zielinski, 2017). Tällöin tallennustiheydeksi kuutiomillimetriä kohden saadaan $215\text{ Pb} \times 4,29 = 922$ petabittiä kuutiomillimetrille, teoreettisen määrän noustessa $200\text{ Eb} \times 4,29 = 858$ eksabittiin kuutiomillimetriä kohti. Nämä arviot on tehty koskien synteettistä DNA:ta. Hieman aikaisemmin tehdyssä tutkimuksessa puolestaan tiheyttä on laskettu käyttäen hyväksi orgaanista DNA:ta, tarkemmin *Bacillus subtilis* - bakteeria jolle laskettu tallennustiheys vastaa myös

aikaisempien tutkimusten 5 petabittiä per gramma - arviota, eli 23 petabittiä per kuutiomillimetri (Cox, 2001).

Kapasiteettimäärää ja tallennustiheyttä on pyritty havainnollistamaan muun muassa esimerkiksi Church ym. (2012) tutkimuksesta. Tässä tutkimuksessa he kopioivat George Churchin noin 54000-sanaisen kirjan, joka sisältää lisäksi kuvia ja JavaScript-ohjelman, 70 miljardia kertaa, pakaten sen DNA-ketjuun. Kaiken edellä mainitun kokonaisuudeksi DNA:han tallennettuna tuli 5,27 megabittiä. Vastaava määrä fyysisenä vastaisi noin 3500 New Yorkin kaupungin kirjastoa, ja digitaalisena noin 46 kappaletta yhden terabitin kiintolevyä. DNA:na kaiken datan koko vastasi hädin tuskin pölyhiukkasta. (Greengard, 2013). Greengard myös huomauttaa, että jo nyt ollaan lähestymässä fyysisten tallennuslaitteiden kapasiteettimaksimia.

Tuloksista voidaan huomata DNA-ketjuihin liittyvät merkittävät kapasiteettihyödyt. Tallennuskapasiteetti on monituhattokertainen verrattuna nykyisiin menetelmiin tallennustiheyden ollessa huomattavasti suurempi kuin missään muussa tunnistetussa menetelmässä.

4.2 Säilyvyys

Toinen eriteltävä mahdollisuus liittyy datan säilyvyyteen. Tarkastellaan edellisen alaluvun tapaan ensin nykyisiä metodeja ja niiden kykyä säilyttää dataa. Tämän jälkeen tehdään sama DNA-ketjulle, jonka jälkeen määritellään lopputulema. Pitkän säilyvyyden hyötynä on muun muassa se, ettei dataa tarvitse siirrellä paikasta A paikkaan B, kun tallennuslaite ikääntyy.

Otetaan jälleen ensiksi huomioon optiset tallennusmenetelmät, eli CD-levyt, DVD-levyt sekä Blu-ray - levyt. Optisten levyjen käyttöikä on rajana pidetään hetkeä, jolloin kaikki levyllä oleva materiaali ei ole enää saatavilla virheettömästi (Slattery, Lu, Zheng, Byers & Tang, 2004). Tutkimuksella on saatu selville, että CD-levyn käyttöikä on noin 45 vuotta niin sanotuissa normaaleissa olosuhteissa (Library of Congress & National Institute of Standards and Technology, 2007). Altistettaessa kuumuudelle ja kosteudelle datan säilyvyys ja levyn käyttöikä laskee huomattavasti. Samassa tutkimuksessa käsiteltiin myös DVD-levyjä. Näiden kohdalla tulokset olivat hyvin vastaavan kaltaiset kuin CD-levyjenkin kohdalla (Library of Congress & National Institute of Standards and Technology, 2007). Vastaavia tuloksia ollaan saavutettu myös muissa tutkimuksissa, joissa on osoitettu, että CD- ja DVD-levyjen varastointiaika on noin 20-50 vuotta, ja oikeanlaisella safiiripäällysteellä jopa 1000 vuotta (Petrov, Kryuchyn & Gorbov, 2011). Blu-ray - levyistä tehdyssä tutkimuksessa on puolestaan saatu selville, että Blu-ray - levy on kykeneväinen säilömään dataa jopa 500 - 2000 vuotta (Miyagawa, Kitaura, Takahashi, Doi, Habuta, Furumiya, Nishiuchi & Yamada, 2006). Tutkijat asettivat levyjä erilaisiin olosuhteisiin simuloiviin tiloihin, joissa pystyttiin kiihdyttämään säilymisprosessia. Oikeanlaisissa olosuhteissa, jotka tutkimuksessa vastasivat Montrealin kaupungin ilmasto-olosuhteita, Blu-ray - levyjen säilymisajaksi saatiin yli 2000 vuotta (Miyagawa ym, 2006).

Käsitellään seuraavaksi magneettinauhaa. Magneettinauhan datan säilytämiskyvystä tehdyssä tutkimuksessa säilymistä heikentäviksi tekijöiksi tutkijat olivat esitelleet magnetismin vähenemisen, päällyskerroksen orgaanisten materiaalien hajoamisen sekä itse nauhan vakauden (Katayama ym., 2016). Ennen itse tutkimusta, tutkijat arvioivat magneettinauhojen kestävän yli 15 vuotta. Tutkijat kiihdyttivät kulumista asettamalla nauhat koviin lämpötiloihin sekä kosteuksiin. Tutkimuksissa saatiin selville, että kuumat ja kosteat olosuhteet saattavat jopa kymmenkertaistaa magneettinauhan ikääntymisen (Katayama ym., 2016). Lopputulemaksi tutkijat saivat, että magneettinauhat ovat kykeneväisiä säilyttämään dataa toimisto-olosuhteissa varmuudella yli 10 vuotta. Vastaavan kaltaisiin tuloksiin ollaan päästy myös toisessa tutkimuksessa, jossa magneettinauhan datansäilytysiäksi arvioitiin 10-20 vuotta (Rothenberg, 1995). Myös hieman toisen kaltaista informaatiota on löydettävissä. Van Bogart (1995) kirjassaan puhuu magneettinauhoista ja niiden kestävydestä. Hän allekirjoittaa aikaisemmin mainittujen tutkimusten tuloksen mainitsemalla tavallisten magneettinauhojen kestävän noin 10-30 vuotta. Samalla hän kuitenkin mainitsee myös magneettinauhat, joilla olisi jopa 100 vuoden säilymisäika. Toisaalta teoksessa myös puhutaan lyhyen säilymisajan puolesta. Van Bogart (1995) mainitsee, ettei pitkällä säilymisellä ole merkitystä teknologian kuitenkin kehittyessä huomattavasti samassa ajassa.

Tarkastellaan seuraavaksi Flash-muistia sekä HDD- ja SSD - kiintolevyjä. Kiintolevyt ovat tällä hetkellä kaikista yleisin tiedontallennusmenetelmä. Flash-muistista tehdyssä tutkimuksessa arvioidaan, ettei Flash-muisti kykene säilyttämään dataa kuin korkeintaan vuoden (Grupp, Caulfield, Coburn, Swanson, Yaakobi, Siegel & Wolf, 2009). Toisaalta on mainittu myös 100 vuoden säilymisajoista (Atmel, 2015). HDD- ja SSD - kiintolevyjen käyttöikää arvioidessa SSD-kiintolevyille käyttöikä on laskettavissa olemassa olevalla kaavalla. Kaavassa kiintolevyn kapasiteetti, uudelleenkirjoitusmäärä ja käytettävissä olevien uudelleenkirjoitusten prosentuaalisen osuuden tulo jaetaan päivittäisen käytön, kapasiteettitekijän sekä vuodessa olevien päivien tulolla (Olson & Langlois, 2008). SSD-levyn käyttöiästä on tehty myös tutkimusta. Kyseisessä tutkimuksessa 32 GB kokoiselle SSD-levylle simuloitiin eriasteisia päivittäisiä käyttökerroja ottamalla huomioon päivittäisen datan käytön määrä. Testausarvon ollessa suurimmillaan (320 GB / päivä), SSD-levyn käyttöiäksi saatiin vajaa 24 vuotta, ja testausarvon ollessa pienimmillään (3,2 GB / päivä), SSD-levyn käyttöiäksi saatiin yli 2300 vuotta. (Olson & Langlois, 2008). HDD-levyjen kohdalla mitään valmiiksi luotua kaavaa ei ole kehitetty. Tutkimuksissa on kuitenkin arvioitu, että on haastavaa saada digitaaliset tiedontallennusmenetelmät säilömään dataa yli 50 vuotta (Grass ym., 2015).

Otetaan sitten huomioon DNA-ketju. Tutkijat ovat myös onnistuneet eristämään erään bakteerilajin DNA:ta yli 250 miljoonaa vuotta vanhoista fossiileista (Vreeland, Rosenzweig & Powers, 2000). Näin ollen jonkinlainen pohja datan säilymiselle on olemassa. Useat tutkimukset arvioivat DNA-ketjun säilyttävän datan erittäin pitkiä aikoja. Eräässä tutkimuksessa säilymisajaksi arvioidaan sadoistatuhansista miljooniin vuosiin (Cox, 2001). Tässä tutkimuksessa arvioi-

daan myös parhaaksi säilytyspaikaksi *Bacillus subtilis* - bakteerin itiöitä, sillä kyseiset bakteerit ovat kykeneväisiä elämään myös hyvin raadollisissa olosuhteissa (Cox, 2001). Toisaalta myös *Deinococcus radiodurans* - bakteeria on ehdotettu isännäksi sen kyvyn selvitä hyvin haastavissa olosuhteissa vuoksi (Wong, Wong & Foote, 2003). Synteettisen DNA:n suhteen tutkijat ovat arvioineet, että kapseloitaessa silikoniin, DNA voi säilyä optimiolosuhteissa jopa 2 miljoonaa vuotta (Grass, Heckel, Puddu & Paunescu, 2015). Hieman varovaisemmissakin arvioissa on puhuttu vähintään tuhansista vuosista, ollen silti konservatiivisia arvioiden suhteen (Goldman ym., 2003).

DNA:n mahdollisuudet toimia pitkäaikaisena datan säilyttäjänä ovat huomattavat nykyisiin menetelmiin verrattuna. Optiset asemat ovat kykeneväisiä säilömään dataa noin 50 vuotta, Blu-ray - levyt 500 - 2000 vuoteen. Magneettinauhoilla tämä olisi vain kymmenestä 30 vuoteen. Kiintolevyjen suhteen yli 50 vuoden säilymisaikaa pidetään haasteellisena. Näihin suhteutettuna tutkijoiden arviot yli miljoonan vuoden säilymisajasta ovat huomattavia. Datan siirtämiselle laitteelta toiselle ei tällöin olisi tarvetta. On huomautettu, että DNA olisi juuri parhaimmillaan datan pitkäaikaisen varastoinnin menetelmänä (Cox, 2001).

4.3 Virheettömyys

Edellä mainittuun liittyy myös DNA-ketjujen kolmas mahdollisuus datan tallennusmenetelmänä nykyisiin tiedontallennusmenetelmiin verrattuna. Pitkästä säilymisajasta ei välttämättä ole paljoa iloa, jollei data pysy mahdollisimman virheettömänä. Tähän ongelmaan DNA-ketjuilla pyritään luomaan vastaus. Tarkastellaan kuitenkin jälleen ensin muita tiedontallennusmetodeja, ennen DNA:n lähempää tarkastelua ja yhteenvetoa.

Aloitetaan jälleen kerran optisista tiedontallennusmenetelmistä, eli CD-levyistä, DVD-levyistä sekä Blu-ray - levyistä. CD- ja DVD - levyille yhteistä on herkkyyys lämpötilavaihteluille, kosteusvaihteluille sekä valolle altistuminen. Nämä tekijät ovat omiaan heikentämään optisten tallennusmenetelmien luotavuutta ja virheensietokykyä. CD-levyjen virheensietokykyä on tutkittu aikaisemmin. Tutkimuksessa pyrittiin luomaan olosuhteet, jotka vastaisivat levyjen luonnollista ikääntymisprosessia. Tulosten esittämisessä on käytetty hyväksi BLER (Block Error Rate) -asteikkoa, jolla optisten levyjen virheherkkyyttä kyetään mittaamaan. Kyseisessä menetelmässä lasketaan levyn datalohkoissa esiintyviä pieniäkin virheellisyyksiä. Sataa CD-levyä testattiin seitsemän vuoden ajan huomioiden niissä syntyvät virheet. Tutkimuksen tuloksina saatiin ajan myötä kasvava virheellisten levyjen määrä. Levyissä huomattiin myös merkittävä määrä kulumaa. (Shanani, Manns & Youket, 2003; Slattery ym., 2004).

DVD-levyjen virheellisyyttä ei kyetä mittaamaan samalla asteikolla kuin CD-levyjä, vaan käytössä on PIE (Parity Inner Errors) ja POE (Parity Outer Errors) - muuttujat. Kuten aikaisemmin on jo mainittu, data jakautuu DVD-levyllä kahteen jonoon. Näistä PIE - virheet ovat virheitä sisemmässä jonossa ja

POE - virheet virheitä ulommassa jonossa. DVD-levyn on mahdollista toimia, vaikka siinä olisi PIE-virheitä. POE-virheellisyyksistä DVD-levy ei kuitenkaan kykene selviytymään. Altistettaessa DVD-levyjä vastaavan kaltaisille olosuhteille kuin CD-levyjäkin, huomattiin virheiden määrässä kasvava trendi. Tutkijat kuitenkin huomauttavat, että virheherkkyys voi vaihdella levyn valmistajasta riippuen. (Slattery ym., 2004). Blu-ray - levyjen kohdalla on saatu myös vastaavanlaisia tuloksia. Altistettaessa korkeille lämpötiloille tai kosteudelle, on levyjen virheellisyys kasvanut (Trivellin, Meneghini, Meneghesso, Zanoni, Orita, Yuri, Tanaka & Ueda, 2009). Myös käytön osoitettiin kasvattavan virheherkkyttä. On myös huomautettu, että naarmut, pölyhiukkaset ja sormenjäljet voivat aiheuttaa optisille tiedontallennusmenetelmille virheellisyyksiä, näistä herkimmin Blu-ray - levyille niiden pakatessa dataa muita optisia menetelmiä tiheämmin (Blu-ray Disc Founders, 2004).

Magneettinauhojen kohdalla luotettavuus mainittiin aiemmin tässä luvussa yhtenä syynä sen käyttöön (Katayama ym., 2016). Magneettinauhojen stabiiliteetiksi on arvioitu yli 10 vuotta, joka 10-30 - vuoden käyttöikä nähdessä on hyvin. Toisaalta säilöminen korkeammassa lämpötiloissa voi nopeuttaa kulumista jopa kymmenkertaiseksi, jolloin myös virheherkkyys kasvaa huomattavasti. (Katayama ym., 2016).

Kiinnitetään seuraavaksi huomio Flash-muistiin sekä HDD- ja SSD - kiintolevyihin. Flash-muistin on mainittu olevan melko virheherkkä ja vaikeasti hallittavissa oleva tallennusmetodi (Cappelletti, 1998). Myös useat ehdotukset parantaa Flash-muistin luotettavuutta ovat puolestaan vaikuttaneet heikentävästi sen tehokkuuteen sekä kapasiteettiin (Chang, Huang, Hsu, Lee, Kuo & Du, 2013). Suurimmaksi yksittäiseksi ongelmaksi Flash-muistin kohdalla nostetaan muistin tyhjennykseen ja kestävyysliittävät ongelmat, sekä lukuhäiriöt, joissa yksi tai useampi bitti muuttuu lukuoperaation aikana. Datan katoamista voi myös tapahtua varauksien muuttuessa tai useiden poisto/ohjelmointi - operaatioiden aikana. (Micron, 2006; Olson & Langlois, 2008). SSD-kiintolevyillä ongelmat ovat vastaavan kaltaisia kuin Flash-muistilla, sillä SSD-kiintolevyt pohjaavat toimintansa Flash-muistiin (Olson & Langlois, 2008).

HDD-kiintolevyt on puolestaan mainittu luotettavuutensa puolesta yhtenä nykyisten tiedontallennusmenetelmien heikoista lenkeistä, ellei jopa heikoimpana lenkinä (Elerath, 2009). Artikkelissa HDD-kiintolevyjen ongelmat jaetaan kahteen osaan: niihin, jotka hajottavat koko kiintolevyn eli ns. operationaaliset virheet, ja niihin, jotka vain korruptoivat kiintolevyn dataa, eli ns. latentssivirheet. Operationaalisia virheitä voivat olla esimerkiksi se, ettei kiintolevylle voi kirjoittaa tai se ettei onnistuneesti kirjoitettua dataa kyetä enää lukemaan (Elerath, 2009). Nämä voivat johtua mm. luku- ja kirjoituspäiden huonosta asemoinnista, joka puolestaan voi johtua huonosta servodatasta, jota ei kyetä muuttamaan, sillä se on kirjoitettu kiintolevyn valmistusprosessin aikana. Servodata puolestaan vastaa siitä, että kirjoitus- ja lukupää tietävät, onko kyseessä lukuoperaatio, kirjoitusoperaatio vai etsintäoperaatio (Elerath, 2009). HDD-levyjen muistikiekat eivät myöskään välttämättä ole täysin pyöreitä, joka voi aiheuttaa kirjoitus/lukupään lukkiutumisen tiettyyn kohtaan, joka voi jälleen

aiheuttaa ongelmia ja virheellisyyksiä (Elerath, 2009). Latenssivirheitä puolestaan ovat virheet, joissa data ei kirjoitu tai katoaa onnistuneen kirjoitusoperaation jälkeen. Nämä voivat aiheutua esimerkiksi siitä, että kirjoitus- ja lukupäiden etäisyys itse levyyn on pienempi kuin nolla, tai vaihtoehtoisesti liian suuri, jolloin magneettisuutta päiden ja levyn välille ei pääse syntymään kunnolla, eivätkä luku ja kirjoitusoperaatiot onnistu virheettömästi (Storm, Lee, Tyndall & Khurshudov, 2007; Elerath, 2009). Lukupäiden etäisyyksiin saattaa vaikuttaa pelkästään epätasaisella alustalla kävelemisen aiheuttama värinä (Elerath, 2009). Toinen mahdollinen latenssivirhe on kiintolevyn liikkuvien osien voiteluaineiden kerääntyminen, joka saattaa siirtää luku- ja kirjoituspäät liian kauas levystä, jolloin dataa ei jälleen kyetä kirjoittamaan tai lukemaan virheettömästi (Elerath, 2009). Lisäksi hetkelliset korkeat lämpötilat kykenevät pyyhkimään dataa kiintolevyltä (Elerath, 2009).

Otetaan taas käsittelyyn DNA-ketjut. Kuten muutkaan aikaisemmin esitellyt tiedontallennusmenetelmät, ei datan DNA-ketjuihin tallentaminen myöskään ole täysin virheetöntä. Toisaalta se on huomattavasti vähemmän virheherkkää muihin metodeihin verrattuna, vaikkakin altistettaessa suojaamattomana elinolosuhteille se on helposti tuhottavissa (Wong ym., 2003). Varhaiselakin teknologialla on saavutettu kirjoittaessa riittävän pieniä virhemarginaaleja, etteivät virheet vaikuta datan luettavuuteen (Cox, 2001). Luonnollisen DNA:n suhteen on tosin huomautettu, että isäntäolento saattaa tulkita koodattua dataa väärin, jolloin tämä voi aiheuttaa virheellisyyksiä sekä datassa että sen kantajassa johtaen pahimmillaan kantajan kuolemaan (Wong ym., 2003). Synteettisen DNA-ketjun kanssa vastaavan kaltaisiin ongelmiin ei kuitenkaan päädytä. Synteettisen DNA:n puolesta puhuu myös se, että koodattaessa DNA-ketjuun 5,27 miljoonaa bittiä, vain 10 näistä biteistä oli virheellisiä (Chuch ym., 2012). Tutkimuksissa ollaan saavutettu myös täydellisiä virheettömyyksiä kirjoitus/luku - prosessissa (Goldman ym., 2013). Tutkimuksessa tallennettiin noin 750 kilobitin edestä tietokoneella olleita tiedostoja ~153000 DNA-ketjuun. Tämän jälkeen ketjut syntetisoitiin ja niistä muodostettiin noin 12 miljoonaa kopiota. Tämän jälkeen ketjut dekodattiin onnistuen saamaan neljä alkuperäisestä viidestä ketjusta ilman esteitä. Yhden ketjun pienet virheet pystyttiin korjaamaan algoritmein, jonka jälkeen dataa luettaessa kaikki alkuperäinen data oli tallessa. (Goldman ym., 2013). Virheettömiä tuloksia on saavutettu myös kyseisen tutkimuksen jälkeen käyttäen hyväksi virheidenkorjausalgoritmeja (Grass, ym., 2015; Blawat ym., 2016; Erlich & Zielinski, 2017). Näiden perusteella puheet virheettömyyden saavuttamisen mahdollisuudesta DNA-ketjujen kohdalla saatavat päätyä uuteen valoon.

DNA-ketjun hyödyt tiedontallennusmenetelmänä ovat selkeät myös virheettömyyden osalta. Käytännön tutkimuksissa ollaan saavutettu virheettömiä tuloksia (Grass, ym., 2015; Blawat ym., 2016; Erlich & Zielinski, 2017). Tähän eivät muut tällä hetkellä käytössä olevat tiedontallennusmenetelmät kykene välttämättä edes virheidenkorjauskoodien avulla.

5 Ongelmat datan tallentamisessa DNA-ketjuihin nykyisiin tiedontallennusmenetelmiin nähden

Mikään teknologia ei ole koskaan täydellistä muihin käytössä oleviin teknologioihin nähden. Tämä pätee myös tarkasteltaessa DNA-ketjujen sopivuutta datan tallennusmenetelmänä. Tässä luvussa otetaan puolestaan kantaa siihen, mitä haasteita datan tallentamisessa DNA-ketjuihin on verraten näitä jälleen nykyään käytössä oleviin datan tallennusmenetelmiin. Näitä ovat muun muassa menetelmän kustannukset sekä lukunopeuden hitaus.

5.1 Kustannukset

Ensimmäinen huomioon otettava asia DNA-ketjuihin tallentamiseen liittyvistä ongelmista on menetelmän hintavuus. Tarkastellaan hintavuutta käyttämällä suurena \$/Mb, joka kertoo näin, kuinka paljon yhden megabitin tallentaminen ja lukeminen maksaa Yhdysvaltain dollareissa. Tämä kertoo paljon muun muassa siitä, kuinka käytännöllinen käytettävä ratkaisu on. Verrataan tätä aspektia jälleen nykyisin käytössä oleviin tiedontallennusmenetelmiin. Kustannukset on niiden havainnollistamiseksi asetettu taulukkoon 2.

TAULUKKO 2 Tallennusmenetelmien kustannukset

Tallennusmetodi	Kustannukset (\$/Mb)
CD-levy	0,0000012
DVD-levy	0,000005
Blu-Ray - levy	0,00003
Magneettinauha	0,00003
Flash-muisti	0,0007
HDD-kiintolevy	0,00007
SSD-kiintolevy	0,003
DNA-ketju	~3500 - 13000

Aloitetaan tarkastelu jälleen optisista tiedontallennusmenetelmistä. arvioiden mukaan CD-levylle dataa tallennettaessa, jäävät kustannukset hyvin mataliksi. Viime vuosituhanen lopulla tehdyssä tutkimuksessa arvioitiin, että 100 megatavun varastointi CD-levyihin maksaisi alle 0,01 Yhdysvaltain dollaria (Porter, Kiesler & Stedfast, 1990). Toisin sanoen tämä vastaisi 0,01 Yhdysvaltain dollarin hintaa per 800 megabittiä. Tämä tarkoittaa valitulla asteikolla 0,0000012 \$/Mb. Eräässä verkossa olevassa artikkelissa viime vuosilta on arvioitu eri muistilaitteiden kustannuksia per gigabitti. Tämän mukaan CD-levyille tallentamisen hinta olisi 0,25 Yhdysvaltain dollaria gigatavulta eli 0,03 Yhdysvaltain dollaria gigabitiltä, joka tämän artikkelin asteikon mukaisesti olisi 0,00003 \$/Mb, eli enemmän kuin edellisen artikkelin arvio (Blank Media Printing, 2016). Näistä ensimmäisellä artikkelilla on kuitenkin tieteellistä pohjaa tukena, joten tulos on luotettavampi. Samalla sivustolla DVD-levyille kustannuksiksi arvioidaan 0,04 Yhdysvaltain dollaria gigatavulta (0,005 Yhdysvaltain dollaria gigabitiltä), eli noin 0,000005 \$/Mb (Blank Media Printing, 2016). Vastaavan kaltaiseen arvioon ollaan päädytty myös toisessa artikkelissa (Menga, 2011). Blu-ray - levyjen kohdalla tehdyssä tutkimuksessa tallentamisen hinnaksi arvioitiin 0,03 Yhdysvaltain dollaria gigabitiltä, joka tarkoittaisi noin 0,00003 \$/Mb (Wan, Cao & Xie, 2014). Blank Media Printingin (2016) artikkelissa hinnaksi arvioitiin puolestaan 0,04 Yhdysvaltain dollaria gigatavulta (0,005 Yhdysvaltain dollaria gigabitiltä), eli 0,000005 \$/Mb. Ensimmäinen näistä arvioista on kuitenkin jälleen tieteellisesti saavutettu ja täten luotettavampi.

Otetaan seuraavaksi tarkasteluun magneettinauhat. Wan ym. (2014) ovat tutkimuksessaan arvioineet myös magneettinauhan tallennuskustannuksia, arvioiden näiden olevan noin 0,03 Yhdysvaltain dollaria gigabitiltä, eli 0,00003 \$/Mb. Vastaavan tulokseen ollaan päädytty myös toisessa artikkelissa (Fontana ym., 2012). Eräässä artikkelissa on tutkittu useampaa magneettinauhaa ja niiden kustannuksia. Tässä magneettinauhoihin tallentamisen kustannukset vaihtelivat 0,001 \$/megatavulta eli 0,00012 \$/Mb aina 0,025 \$/megatavulta eli 0,003125 \$/Mb (Drapeu & Katz, 1993). Vaihtelut johtuvat tutkimuksessa käytettyjen laitteiden hintojen eroavaisuuksista sekä Drapeun ja Katzin (1993) tutkimuksen mahdollisesti vanhanaikaisista tiedosta.

Seuraavana vuorossa ovat Flash-muisti sekä HDD- ja SSD - kiintolevyt. Fontana ym. (2012) tutkivat artikkelissaan myös HDD- ja SSD - kiintolevyjä. HDD - kiintolevyjen tallennuskustannuksiksi on arvioitu 0,07 Yhdysvaltain dollaria gigabitiltä eli 0,00007 \$/Mb. SSD-kiintolevyillä vastaava arvo on kolme Yhdysvaltain dollaria gigabitiltä eli 0,003 \$/Mb. Toisaalta tutkijat arvioivat tämän lukeman laskevan tulevaisuudessa 0,00019 \$/Mb (Fontana ym., 2012). HDD - kiintolevyjen kohdalla vastaavan kaltaisia tuloksia on saavutettu myös toisessa tutkimuksessa (Wan ym., 2014). He myös arvioivat Flash-muistin kustannusten olevan noin kymmenkertaiset verrattuna HDD-kiintolevyihin.

Verrattuna nykyään käytössä olevien tiedontallennusmenetelmien synnyttämiin kustannuksiin, ovat ne varsin pieniä verrattaessa niitä DNA-ketjuihin tallentamisen synnyttämiin kuluihin. On arvioitu, että kirjoitusoperaatio DNA-

ketjuun maksaa noin 12400 \$/Mb ja lukuoperaatiokin noin 220 \$/Mb (Goldman ym., 2013). Erlich & Zielinski (2017) arvioivat kustannuksiksi noin 3500 \$/Mb. Tallentamisen hinta on pudonnut vuosittain noin 12-kertaista vauhtia (Church ym., 2012). On kuitenkin esitetty arvioita, että DNA-ketjuihin tallentaminen olisi kustannustehokasta jopa alle 50-vuoden arkistojen ylläpitämiseen lähivuosien aikana, ja käytännöllistä seuraavan 50 vuoden aikana (Goldman ym., 2013). Myös Blawat ym. (2016) uskovat datan tallentamisen DNA-ketjuihin olevan taloudellisesti järkevää jo seuraavan 5-10 vuoden aikana.

Kuten näistä tuloksista nähdään, on DNA:lla vielä paljon parannettavaa kustannustehokkuuden suhteen verrattuna nykyään käytössä oleviin tiedontallennusmenetelmiin, jotka ovat satojatuhansia kertoja halvempia per megabitti. Toisaalta teknologian kehitys voi mahdollistaa tämän yleistymisen kuten Church ym. (2012) huomautus antaa uskoa, kuten myös huomio tallentamisen 9000 \$/Mb halpenemisestä neljässä vuodessa (Goldman ym., 2013; Erlich & Zielinski, 2017).

5.2 Lukunopeus

Yleisesti data halutaan saada nopeasti esille. Tämän vuoksi nykyiset tallennusmenetelmät ovat kehittyneet jatkuvasti nopeammiksi. Tämän valossa tulevaisuuden teknologian tulisi olla ainakin vähintään yhtä nopeaa ollakseen yleisesti käyttökelpoista. Tässä alaluvussa tarkastellaan toista DNA-ketjuihin tallentamiseen liittyvää haastetta verraten sitä nykyään käytössä oleviin tiedontallennusmenetelmiin. Lukunopeudet on esitetty niiden havainnollistamiseksi taulukossa 3.

TAULUKKO 3 Tallennusmenetelmien lukunopeudet

Tallennusmetodi	Lukunopeus
CD-levy	19,2 Mb/s
DVD-levy	76,8 Mb/s
Blu-ray - levy	36 Mb/s - 72 Mb/s
Magneettinauha	240 Mb/s - 2880 Mb/s
Flash-muisti	245,6 Mb/s - 400 Mb/s
HDD-kiintolevy	~1280 Mb/s
SSD-kiintolevy	< 1600 Mb /s - ~4480 Mb/s
DNA-ketju	0,01 Mb/s

Aloitetaan tarkastelu optisista tiedontallennusmenetelmistä. On arvioitu, että CD-levyillä lukunopeudet ovat noin 2,4 megatavua eli 19,2 megabittiä sekunnissa (Chervenak, 1998). Samassa tutkimuksessa DVD-levyille arvioitu lukunopeus on 9,6 megatavua eli 76,8 megabittiä sekunnissa. Blu-ray - levyillä lukunopeudet ovat yllättäen hieman hitaammat kuin DVD-levyillä. Blu-ray - levyjen spesifikaation mukaan lukunopeus vaihtelee 36 megabittiä sekunnissa

ja 72 Mb/s välillä riippuen, onko kyseessä versio 1.0 vai 2.0 (Blu-ray Disc Founders, 2004). Seuraavaksi tarkasteltavilla magneettinauhoilla lukunopeudet ovat jälleen hieman suurempaa luokkaa. Useaa eri magneettinauhaa testanneessa tutkimuksessa näiden lukunopeudet vaihtelivat aina 30 megatavusta eli 240 megabitistä sekunnissa jopa yli 120 megatavuun eli 960 megabittiin sekunnissa (Dee, 2008). IBM on puolestaan ilmoittanut oman magneettinauhateknologiansa yltävän jopa 360 megatavun eli 2880 Mb/s lukunopeuksiin (IBM, 2014).

Otetaan seuraavaksi huomioon Flash-muisti, HDD-kiintolevyt sekä SSD-kiintolevyt. Flash-muistin lukunopeudeksi on arvioitu 30.7 megatavua sekunnissa, joka tarkoittaa 245,6 megabittiä sekunnissa (Caulfield, Grupp & Swanson, 2009). Toisessa artikkelissa kuitenkin mainitaan Flash-muistiin liittyvä pursekeellisuus, jolla voidaan hetkellisesti tehostaa lukuoperaatioita, jolloin lukunopeus voi nousta jopa 50 megatavuun eli 400 megabittiin sekunnissa (Bez, Camerlenghi, Modelli & Visconti, 2003). Tarkasteltaessa HDD-kiintolevyjä kiintolevyvalmistaja SanDisk mainitsee HDD-kiintolevyn lukunopeudeksi 156 megatavua eli 1248 megabittiä sekunnissa (SanDisk, 2016). Tarkasteltaessa toisen kiintolevyvalmistaja Seagaten HDD-kiintolevyjä, lukunopeudet ovat noin 160 megatavun eli 1280 megabittiä sekunnissa luokkaa (Seagate, 2018). SSD-levyjen lukunopeudesta mainitaan muun muassa tutkimuksessa, jossa lukunopeudeksi arvioitiin noin 200 megatavua sekunnissa, joka vastaa 1600 megabittiä sekunnissa (Chen, Lee & Zhang, 2011). Toisaalta SSD-kiintolevyjen valmistajien sivuilta tarkastellessa esimerkiksi Samsungin SSD 860 PRO 2.5" SATA III 4TB -kiintolevy kykenee 560 megatavun eli 4480 megabittiä sekunnissa lukunopeuksiin (Samsung, 2018). SSD-levyt ovat siis huomattavasti HDD-kiintolevyjä nopeampia.

Otetaan seuraavaksi käsittelyyn DNA-ketjut. Esimerkiksi Church ym. (2012) artikkelissaan toteavat, että tämänhetkiset kirjoitus- ja lukunopeudet DNA-ketjun suhteen ovat hyvin riittämättömät muuhun kuin hyvin pitkään säilytettäviin arkistoihin. Cox (2001) puolestaan mainitsee lukunopeuden yhtenä suurimmista ongelmista. Yhden bitin lukemisen DNA:sta on arvioitu vievän 100 mikrosekuntia (Zhirnov, Zagedan, Sandhu, Church & Hughes, 2016). Tämä tarkoittaisi, että yhden megabitin lukemiseen kuluisi 100 sekuntia, jolloin DNA-ketjun lukunopeudeksi saadaan 10 kilobittiä sekunnissa.

Tarkastellessa äsken esitettyjä lukunopeuksia, huomataan, että DNA-ketjuista datan lukeminen on huomattavasti hitaampaa nykyisiin tiedontallennusmenetelmiin nähden. Ero nopeimpiin menetelmiin verrattuna on yli 10000-kertainen näiden menetelmien eduksi.

6 Yhteenveto

Tuotamme jatkuvasti valtaiset määrät uutta dataa, jolle tarvitaan toimiva ratkaisu sen tallentamiseksi. Vaikka käytössämme on jo laaja valikoima tiedontallennusmenetelmiä, vain harvat niistä sopivat datan pitkäaikaiseen tallentamiseen. Moneen näistä liittyy datan säilyvyyteen ja virheellisyyteen liittyviä ongelmia, jotka pakottavat siirtämään dataa tallennuslaitteelta A tallennuslaitteelle B, joka myös on omiaan aiheuttamaan mahdollisia ongelmia tiedon säilyvyyden kannalta. Tähän syntyneeseen ongelmaan on ehdotettu ratkaisuksi DNA-ketjujen valjastamista tiedontallennukseen. DNA-ketju on luontainen datan varasto, sillä se sisältää jokaisen elollisen olennon perimätiedon ja datan. Tämän tutkielman tarkoituksena oli tarkastella sitä, onko DNA-ketjuista tuomaan ratkaisua suurten datamassojen tallentamisen välineeksi. Tähän ongelmaan pyrittiin löytämään selvennystä vertailemalla DNA-ketjujen synnyttämiä mahdollisuuksia ja haasteita suhteessa tällä hetkellä olemassa oleviin tiedontallennusmenetelmiin nähden.

Datan tallentaminen DNA-ketjuihin tarjoaa monia hyötyjä suhteessa nykyään käytössä oleviin tiedontallennusmenetelmiin nähden. Dataa voidaan varastoida huomattavia määriä enemmän ja huomattavasti tiheämmin kuin millään muulla olemassa olevalla tiedontallennusmenetelmällä tänä päivänä käytään. Esimerkiksi kapasiteettinsa suhteen DNA-ketjut päihittävät nykyiset kiintolevyt monisatatuhatkertaisesti tallennustiheyden ollessa vielä huomattavasti moninkertaisempi. DNA-ketjut mahdollistavat myös datan pitkäaikaisen säilyttämisen. Optiset tiedontallennusmenetelmät kykenevät muutamaa kymmeniä vuosiin modernimpien kiintolevyjen jäädessä huomattavasti myös tästä. DNA-ketju voi puolestaan säilyä oikein varastoituna jopa kaksikin miljoonaa vuotta. Kolmantena esille nostettuna mahdollisuutena DNA-ketjuissa on niiden kyky säilyttää dataa virheettää. Monissa viimeisimmissä tutkimuksissa tutkijat ovat saaneet luettua sinne tallennetut tiedostot täysin virheettöminä käyttäen virheenkorjausalgoritmeja hyväkseen. Vaikka myös kaikki nykyisetkin tiedontallennusmenetelmät käyttävät hyväkseen virheenkorjausalgoritmeja, voi esimerkiksi HDD-kiintolevyissä syntyä sellaisia virhetiloja, joita ei pystytä korjaamaan edes virheenkorjausalgoritmien avulla.

Vaikka DNA-ketjut tarjoavatkin huomattavia mahdollisuuksia nykyisiin tiedontallennusmenetelmiin nähden, tässä teknologiassa on myös omat haasteensa. Kun muiden tiedontallennusmenetelmien tallennuskustannukset voidaan laskea yhden Yhdysvaltain dollarin murto-osissa per megabitti, vastaavasti DNA-ketjuilla summa on tuhansia Yhdysvaltain dollareita per megabitti. Toisaalta tutkijat ovat luottavaisia sen suhteen, että jo lähiaikoina DNA-ketjuihin tallentamisesta tulee kustannustehokasta. Lisäksi menetelmän kustannukset ovat laskeneet jopa kahdestoistaosaan edellisistä vuositasolla. Toinen ongelma, mikä DNA-ketjuissa on nykyisiin tiedontallennusmenetelmiin nähden, on sen vaatimaton lukunopeus. Tämän tutkijat ovat nostaneet yhdeksi merkittävimmistä ongelmista DNA-ketjuihin tallentamiseen liittyen. Monet nykyisiä menetelmistä ovat jopa tuhat kertaa nopeampia lukunopeudeltaan DNA-ketjuihin verrattuna.

Tutkielman aihetta ei pyritty käsittelemään syväluotaavasti vaan keskityttiin vertaamaan hyötyjä ja ongelmia suhteessa nykyisiin tiedontallennusmenetelmiin, jotka ovat kaupallisessa käytössä. Tutkielman tekemisen aikana esiin tuli monia kehitysasteella olevia teknologioita, jotka lupaavat korkeaa kapasiteettia ja tehokkuutta. DNA-ketjuihin tallentamisen tutkimista voitaisiin jatkossa ulottaa myös näiden kehitteillä olevien teknologioiden kanssa vertaamiseen. Lisäksi tutkimusta voitaisiin tehdä mahdollisista käytännön sovelluksista, joihin DNA-ketjuja voitaisiin tiedontallennusmenetelmänä käyttää.

LÄHTEET

- Amankwah-Amoah, J. (2016). Competing technologies, competing forces: The rise and fall of the floppy disk, 1971–2010. *Technological Forecastin and Social Change*, 107, 121-129.
- Atmel (2015). 8-Bit AVR Microcontroller ATmega32A Datasheet Complete. Atmel Corporation.
- Bauer, F.L. & Wössner, H. (1972). The “Plankalkül” of Konrad Zuse: A Forerunner of Today’s Programming Languages. *Communications of the ACM*, 15(7), 678-685.
- Bez, R., Camerlenghi, E., Modelli, A. & Visconti, A. (2003). Introduction to Flash Memory. *Proceedings of the IEEE*, 91(4), 489-502.
- Blank Media Printing (2016, 26. tammikuuta). Cost Per Gigabyte of Popular Data Storage - Infographic. Haettu 14.4.2018 osoitteesta : <https://www.blankmediaprinting.com/blog-article/cost-gigabyte-popular-data-storage-types-infographic>
- Blawat, M., Gaedke, K., Hütter, I., Chen, X-M., Turczyk, B., Inverso, S., Pruitt, B. & Church, G. (2016). Forward Error Correction for DNA Data Storage. *Procedia Computer Science*, 80, 1011-1022.
- Blu-ray Disc Founders (2004). Blu-ray Disc Format : 1.A Physical Format Specifications for BD-RE. White Paper.
- Campbell, N., Reece, J., Urry, L., Cain, M., Wasserman, S., Minorsky, P. & Jackson, R. (2015). *Biology: A Global Approach* (10. painos). Essex : Pearson.
- Cappelletti, P. (1998). Flash memory reliability. *Microelectronics Reliability* 38(2), 185-188.
- Caulfield, A.M, Grupp, L.M & Swanson, S. (2009). Gordon: Using Flash Memory to Build Fast, Power-efficient Clusters for Data-intensive Applications. Julkaisussa *Proceedings of the 14th international conference on Architectural support for programming languages and operating systems (ASPLOS’09)* (217-228), Washington D.C., March 7-11 2009.
- Chang, Y-H., Huang, P-C., Hsu, P-H., Lee, L-J., Kuo, T-W. & Du, D.H-C. (2013). Reliability Enhancement of Flash-Memory Storage Systems : An Efficient Version-Based Design. *IEEE Transaction on Computers*, 62(12), 2503-2515.
- Chen, F., Lee, R. & Zhang, X. (2011). Essential roles of exploiting internal parallelism of flash memory based solid state drives in high-speed data processing. Julkaisussa *2011 IEEE 17th International Symposium on High Performance Computer Architecture* (266-277), February 2011.
- Chervenak, A. (1998). Challenges for tertiary storage in multimedia servers. *Parallel Computing*, 24(1), 157-176.
- Christman, E. (2018 2. helmikuuta). Best Buy to Pull CDs, Target Threatens to Pay Labels for CDs Only When Customers Buy Them. Haettu 16.4.2018 osoitteesta : <https://www.billboard.com/articles/business/8097929/best-buy-to-pull-cds-target-threatens-to-pay-labels-for-cds-only-when>

- Church, G., Gao, Y. & Kosuri, S. (2012). Next-Generation Digital Information Storage in DNA. *Science*, 337(6102), 1628.
- Copeland, B.J. (2004). Colossus: Its Origins and Originators. *IEEE Annals of the History of Computing*, 26(4), 38-45.
- Coughlin, T. (2017). A Timeline for Flash Memory History [The Art of Storage]. *IEEE Consumer Electronics Magazine*, 6(1), 126-133.
- Cox, J. (2001). Long-term data storage in DNA. *Trends in Biotechnology*, 19(7), 247-250.
- De Silva, P. Y., & Ganegoda, G. U. (2016). New Trends of Digital Data Storage in DNA. *BioMed Research International*, 2016.
- Dee, R. (2008). Magnetic Tape for Data Storage: An Enduring Technology. *Proceedings of the IEEE*, 96(11), 1775-1785.
- Drapeu, A. & Katz, R. (1993). Striped tape arrays. *Julkaisussa the 12th IEEE Symposium on Mass Storage Systems*. 26.4. - 29.4.1993, 257-265.
- Elerath, J. (2009). Hard-Disk Drives: The Good, the Bad, and the Ugly. *Communications of the ACM*, 52(6), 38-45.
- Erlich, Y. & Zielinski, D. (2017). DNA Fountain enables a robust and efficient storage architecture. *Science*, 355(6328), 950-954.
- Fontana, R.E., Hezler, S.R. & Decad, G. (2012). Technology Roadmap Comparisons for TAPE, HDD, and NAND Flash: Implications for Data Storage Applications. *IEEE Transaction on Magnetics*, 48(5), 1692-1696.
- Greengard, S. (2013). A New Approach to Information Storage. *Communications of the ACM*, 56(8), 13-15.
- Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E., Sipos, B. & Birney, E. (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, 494, 77-80.
- Gonzales, D., Winkelman, Z., Tran, T., Sanchez, R., Woods, D. & Hollywood, J. (2017). Digital Forensics Compute Cluster: A High Speed Distributed Computing Capability for Digital Forensics. *International Journal of Computer and Information Engineering*, 11(8), 916-923.
- Grass, R., Heckel, R., Puddu, M. & Paunescu, D. (2015). Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes. *Angewandte Chemie International Edition*, 54(8), 2552-2555.
- Gray, J. (1996). Evolution of Data Management. *Computer*, 29(10), 38-46.
- Grupp, L.M., Caulfield, A.M, Coburn, J., Swanson, S., Yaakobi, E., Siegel, P.H. & Wolf, J.K. (2009). Characterizing flash memory: Anomalies, observations, and applications. *Teoksessa Microarchitecture, 2009. MICRO-42. 42nd Annual IEEE/ACM International Symposium on 2009 42nd Annual IEEE/ACM International Symposium on Microarchitecture (24-33)*, Dec. 2009.
- Hilbert, M. & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60-65.
- IBM (2014). IBM TS1150 tape drives deliver the fastest and largest capacity drive for enterprise archiving and data protection. IBM United States Hardware Announcement, 114-165.
- IBM (2015). The Floppy Disk. Haettu 15.4.2018 osoitteesta : <http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/floppy/>

- Immink, K.A.S. (1996). The Digital Versatile Disc (DVD): System Requirements and Channel Coding. *SMPTE Journal*, 105(8), 483-489.
- Kablau, J.G.F. (1996). The DVD Physical Format. *Teoksessa 1996 Digest of Technical Papers - International Conference on Consumer Electronics*, 5.6 - 7.6.1996.
- Katayama, K., Chinda, Y., Shimizu, O., Mikami, T., Suzuki, M. & Noguchi, H. (2016). Long-Term Stability of Magnetic Tape for Data Storage Under an Accelerated Condition. *IEEE Transaction on Magnetics*, 52(7), 1-4.
- Kilgour, F.G (1997). Origins of Coordinate Searching : Introduction. *Journal of the American Society for Information Science*, 48(4), 340.
- Library of Congress & National Institute of Standards and Technology (2007). NIST/Library of Congress (LC) Optical Disk Longevity Study.
- Menga, R. (2011, 25. tammikuuta). WHICH IS CHEAPER - HARD DRIVE STORAGE OR OPTICAL?. Haettu 14.4.2018 osoitteesta : <https://www.pcmec.com/article/which-is-cheaper-hard-drive-storage-or-optical/>
- Michigami, T. & Tanaka, K. (1998). U.S. Patent No. 6223322B1. Washington D.C : U.S. Patent and Trademark Office.
- Micron (2006). TN-29-17: NAND Flash Design and Use Considerations Introduction. Micron Technology Inc. Technical Note.
- Miyagawa, N. (2014). Overview of Blu-Ray Disc™ recordable/rewritable media technology. *Frontiers of Optoelectronics*, 7(4), 409-424.
- Miyagawa, N., Kitaura, H., Takahashi, K., Doi, Y., Habuta, H., Furumiya, S., Nishiuchi, K. & Yamada, N. (2006). Over 500 Years Lifetime Dual-Layer Blu-Ray Disc Recordable based on Te-O-Pd Recording Material. *Proceedings of SPIE*, 6282.
- Olson, A. & Langlois, D. (2008). Solid State Drives Data Reliability and Lifetime. White Paper : Imation.
- Peek, H.B. (2010). The emergence of the compact disk. *IEEE Communications Magazine*, 48(1).
- Petrov, V., Kryuchyk, A. & Gorbov, I. (2011). High-density optical disks for long term information storage. *Proceedings of SPIE*, 8011.
- Piramanayagam, S.N. & Chong, T.C. (2011). *Developments in Data Storage : materials perspective*. John Wiley & Sons.
- Porter, J.L., Kiesler, J.L. & Stedfast, D.A. (1990). Archiving Online Data on Optical Disk. U.S. Geological Survey. Open-File Report 90-575 : Reston, Virginia.
- Portney, N., Wu, Y., Quezada, L., Lonardi, S. & Ozkan, M. (2008). Length-Based Encoding of Binary Data in DNA. *Langmuir : the ACS journal of surfaces and colloids*, 24(5), 1613-1616.
- Pugh, E.W. & Heide, L. (2013). Early punched card equipment: 1880-1951 [Scanning Our Past]. *Proceedings of the IEEE*, 101(2), 546-552.
- Rothenberg, J. (1995). Ensuring the Longevity of Digital Documents. *Scientific American*, 272(1), 42-47.
- Royon, A., Bourhis, K., Bellec, M., Papon, G., Bousquet, B., Deshayes, Y., Cardinal, T. & Canioni, L. (2010). Silver Clusters Embedded in Glass as a

- Perennial High Capacity Optical Recording Medium. *Advanced Materials*, 22(46), 5282-5286.
- SanDisk (2016). WHICH STORAGE DEVICE WOULD YOU GO TO BATTLE WITH? Solid State vs. Hard Disk. Haettu 14.4.2018 osoitteesta : https://www.sandisk.com/content/dam/sandisk-main/en_us/assets/resources/enterprise/infographics/how-do-ssds-stack-up-against-hdds.pdf
- Samsung (2018). SSD 860 PRO 2.5" SATA III 4TB. Haettu 14.4.2018 osoitteesta : <https://www.samsung.com/us/computing/memory-storage/solid-state-drives/ssd-860-pro-2-5--sata-iii-4tb-mz-76p4t0bw/>
- Seagate (2018). Introducing the Thinnest and Lightest 2.5-Inch Hard Drives. Haettu 14.4.2018 osoitteesta : <https://www.seagate.com/gb/en/internal-hard-drives/hdd/barracuda/>
- Senko, M.E., Altman, E.B, Astrahan, M.M. & Fehder, P.L. (1971). Data structures and accessing in data-base systems, I: Evolution of information systems. *IBM Systems Journal*, 12(1), 30-44.
- Shanani, C., Manns, B. & Youket, M. (2003). Longevity of CD media : research at the library of congress. *Preservation Research and Testing Division, Library of Congress*, Washington D.C.
- Slattery, O., Lu, R., Zheng, J., Byers, F. & Tang, X. (2004). Stability Comparison of Recordable Optical Discs – A Study of Error Rates in Harsh Conditions. *Journal of Research of the National Institute of Standards and Technology*, 109(5), 517-524.
- Stevens, L.D. (1981). The Evolution of Magnetic Storage. *IBM Journal of Research and Development*, 25(5), 663-676.
- Storm, B.D., Lee, S., Tyndall, G.W. & Khurshudov, A. (2007). Hard Disk Drive Reliability Modeling and Failure Prediction. *IEEE Transaction on Magnetics*, 43(9), 3676-3684.
- Trivellin, N., Meneghini, M., Meneghetto, G., Zanoni, E., Orita, K., Yuri, M., Tanaka, T. & Ueda, D. (2009). Reliability analysis of InGaN Blu-Ray laser diode. *Microelectronics Reliability*, 49(9), 1236-1239.
- Ulanoff, L. (2010). The Floppy is Dead: Time to Move Memories to the Cloud. *PC Magazine*, 29(7).
- Van Bogart, J.W.C. (1995). Magnetic Tape Storage and Handling : A Guide for Libraries and Archives. National Media Lab : St. Paul.
- Vreeland, R., Rosenzweig ,W. & Powers, D. (2000). Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal. *Nature*, 407(6806), 897-900.
- Wan, S., Cao, Q. & Xie, S. (2014). Optical storage: an emerging option in long-term digital preservation. *Frontiers of Optoelectronics*, 7(4), 486-492.
- Watson, J. & Crick, F. (1953) . The Structure of DNA. *Cold Spring Harbor Symposia on Quantitative Biology*, 18, 123-131.
- Wong, P., Wong, K-K. & Foote, H. (2003). Organic data memory using the DNA approach. *Communications of the ACM*, 46(1), 95-98.

Zhirnov, V., Zagedan, R.M., Sandhu, G.S., Church, G.M. & Hughes, W.L. (2016).
Nucleic Acid Memory. *Nature Materials*, 15(4), 366.