Dna:sta
lääke dataähkyyn?

Nykyaika tuottaa valtavia määriä tietoa, jonka varastointi muodostaa hankalan pullonkaulan. Ongelman saattaa ratkaista täysin uudenlainen datavarasto: dna.

Jari Koponen

Riisinjyvän kokoinen vaaleanpunainen läikkä ampullin pohjalla sisältää koodattuna sata maailmankirjallisuuden klassikkoteosta.

Tämä noin kahdensadan megatavun kokoinen data on toistaiseksi suurin dna:han kirjoitettu ja siitä virheettömästi luettavissa oleva tiedosto. Viimevuotisen saavutuksen takana olivat Microsoft ja Washingtonin yliopisto.

Ohjelmistojätti ilmoitti hiljattain jatkavansa menetelmän kehittelyä. Tämä kertoo osaltaan, että dna on vakavasti otettava vaihtoehto tulevaisuuden datan tallentamiseen.

New York Genome Center

Tutkijat Yaniv Erlich (vas.) ja Dina Zielinski onnistuivat pakkaamaan dna:han dataa ennätykselliset 215 petatavua grammaa kohden.

Ylivertainen varasto

Dna:lla on muihin tallennusalustoihin nähden ylivoimaisia ominaisuuksia, joita ovat ennen kaikkea sen säilyvyys ja pakkaustiheys.

Eliöissä on geenejä, jotka ovat säilyneet jokseenkin muuttumattomina miljardeja vuosia. Eliöiden ulkopuolella dna:n säilyvyys on oikeissa olosuhteissa lyömätön, mikä näkyy jo fossiileista eristetyistä dna-määristä. Ikijäässä säilyneestä noin 700 000 vuotta vanhasta luusta saatiin riittävästi dna:ta niin, että sen avulla kyettiin rekonstruoimaan eläimen koko genomi.

Viileänä ja kuivana dna säilyy pitkiä ajanjaksoja. Lisäksi asiaan vaikuttaa kopioiden lukumäärä. Mitä suurempi se on, sitä suurempi on todennäköisyys pitkäaikaisesta säilymisestä.

Simulointien perusteella tehdyt laskelmat kertovat, että jos käytetään tuhatta kopiota, saavutetaan tuhannen vuoden säilyvyys 99,999 prosentin varmuudella.

Hurjin kirjallisuudessa esitetty arvio perustuu vanhenemista simuloivaan käytännön kokeeseen. Sen mukaan piihin kapseloitu, Huippuvuorilla sijaitsevan kansainvälisen siemenpankin olosuhteissa eli 18 celsiusasteen pakkasessa varastoitava digitoitu dna-näyte säilyisi yli kaksi miljoonaa vuotta.

Pakkaustiheydessä dna:lle ei ole vertaa. Sen valtteja ovat koodausyksikön eli emäsparin pieni koko – noin puoli nanometriä – ja kolmiulotteisuus.

Lisätiivistys tulee koodausyksikköinä käytetyistä neljästä erilaisesta emäksestä eli adenosiinista, syto-

siinista, guaniinista ja tyminiinista. Niistä kukin voi teoriassa sisältää kaksi bittiä, esimerkiksi: A(00), S(01), G(10) ja T(11). Käytännön syyt tosin pienentävät arvon 1,8 bittiin emäsparia kohden.

Dna:n etuna on myös varastoinnin mitätön energiakulutus, jota voidaan mitata watteina gigatavua kohti. Näin alustan lämpenemisongelmaa ei ole.

Dna-tiedostoa ei voi hakkeroida, ainoastaan varastaa. Nopeasti vanhenevia tallennusteknologioita tulee ja menee. Nauhan, lerpun, korpun ja cd:n mukana häviää samalla niihin tallennettu tieto. Dna ei katoa eikä vanhene.

Monta menetelmää

Digitoidun tiedon muuntamisessa dna:n emäsparien järjestykseksi ja edelleen vastaaviksi dna-ketjuiksi on useampia menetelmiä.

Yhdessä niistä binäärikoodi jaetaan ensin algoritmin avulla satunnaisvalinnalla pienemmiksi palasiksi. Kukin pala sisältää kahtasataa emäsparia vastaavan koodin. Mukana on datan lisäksi virheenkorjauskoodi sekä pätkän alkuun ja loppuun liitetyt, synteesiä varten välttämättömät alukkeet.

Lopullinen binäärikoodi sisältää listan kymmenistätuhansista yksijuosteisista dna-pätkistä eli oligo- ➤➤➤

Scanstockphoto

➤➤➤

nukleotideista (oligoista). Toiminnassa on jo kaupallisia yrityksiä, jotka syntetisoivat koodia vastaavat oligot tekstimuodossa lähetetyn koodin avulla. Tämä vaihe kestää viikon tai pari.

Datan lukemiseen dna:sta käytetään modernia sekvensointimenetelmää eli emäsjärjestyksen määrittämistä. Sekvensoinnin tulokset dekoodataan takaisin biteiksi ja alkuperäiseksi tiedostoksi. Dekoodaus hoituu tavallisella kannettavalla tietokoneella muutamassa minuutissa, minkä jälkeen alkuperäisdata voidaan toistaa virheettömästi.

Menetelmää ovat hyödyntäneet muun muassa New Yorkin genomikeskuksen tutkijat **Yaniv Erlich** ja **Dina Zielinski**. He julkaisivat vuoden alussa *Science*-lehdessä artikkelin, jossa he esittelivät uusia merkittäviä tuloksiaan.

Vaikka kaksikon taltioiman datan määrä, 2,15 megatavua, oli pienempi kuin Microsoftilla, he saavuttivat uuden pakkaustiheysennätyksen eli 1,57 bittiä emäsparia kohden.

Grammalla dna:ta tulos vastaa 215 petatavun (10^{15}) varastointikykyä. Moisesta huipputiheydestä huolimatta datan virheetön kirjoitus ja luenta onnistuivat sataprosenttisesti.

Vertailun vuoksi: suurilla datakeskuksilla on nykyään käytettävissä tallentamiseen enintään 10 teratavun (10^{12}) levyjä toisiinsa kytkettäviksi. Yleisempiä ovat kuitenkin kahdeksan teratavun levyt. Niitä tarvittaisiin lähes 27 000, jotta päästäisiin samaan kapasiteettiin kuin yhdellä dna-grammalla.

Newyorkilaisten toinen merkittävä tulos liittyy datan lukemiseen. Jokainen oligonäytteen luentakerta kuluttaa näytettä. PCR-tekniikalla on kuitenkin mahdollista kopioida oligot ja korvata syntyvä vajaus.

Kaksikko testasi menetelmää monistussarjoilla. Tulokset osoittivat, että vaikka PCR tuottaakin virheitä, data oli silti luettavissa virheettömästi vielä viittä miljoonaa lukukertaa vastaavassa tilanteessa.

Datan hallinta on osa tiedepolitiikkaa

CSC – Tieteen tietotekniikan keskus Oy on valtion ja korkeakoulujen omistama, tutkimuksen tarpeisiin suunniteltujen IT-ratkaisujen erityistehtäväyritys. Yhtiöllä on kaksi konesalia, toinen Espoon Otaniemessä ja toinen Kajaanissa, jossa toimii Suomen tehokkain supertietokone Sisu.

CSC on myös oikea paikka saada tietoa nykyajan datantallennuksen tarpeista ja haasteista.

”Raakadataa säilytetään yleensä mahdollisimman kustannustehokkaasti nauhajärjestelmissä, joista se joudutaan siirtämään kalliimmille levyjärjestelmille laskentaa ja analyysia varten”, aloittaa **Pekka Lehtovuori**, joka toimii CSC:n tutkimuksen palveluista vastaavan yksikön johtajana.

Määrältään nopeimmin kasvavia datamääriä tuottavat hänen mukaansa tätä nykyä genomi- ja ilmastotutkimukset. Tulevaisuudessa laajenee etenkin esineiden internet (IoT) eli laitteiden välinen kommunikointi. Sen myötä dataa alkavat suoltaa muun muassa erilaiset sensorit.

”Nähtäväksi jää, miten ja missä määrin tätä datamäärää käsitellään itse mittalaitteissa ja missä määrin perinteisissä datakeskuksissa.”

Lehtovuoren mukaan yksi ongelma on siinä, että datan tuottaminen ja sen käsittely tapahtuvat tavallisesti eri paikoissa.

”Esimerkiksi genomidataa ei yleensä pystytä käsittelemään samassa paikassa, jossa se tuotetaan. Siten satoja teratavujen ja jopa petatavujen kokoisia tiedostoja joudutaan siirtämään muualle.”

Vastaan tulevat tällöin rajoitukset tiedonsiirtoverkkojen nopeuksissa. Esimerkiksi 50 teratavun siirto sadan megatavun sekuntinopeudella kestää noin viikon.

Hänen mukaansa datan hallinnan suunnitteluun ei aina kiinnitetä riittävästi huomiota. Datan taltiointi voi olla lyhytkestoista tai pitkäaikaissäilytystä. Data voidaan tuhota joko heti käytön tai määräajan jälkeen.

”Kvanttikemian lasku, jonka läpivieminen kesti vuosituhannen vaihteessa viikon, pystytään nykylaittein laskemaan parissa tunnissa. Tulokset, jotka silloin kannatti säilyttää uudelleenkäyttöä varten, saattaa olla helpompaa ja kustannustehokkaampaa laskea nykykoneilla uudelleen”, hän antaa esimerkin.

Pitkäaikaiseen säilytykseen kuuluu ennen kaikkea sellainen data, jota ei voida tuottaa uudestaan. Tällaisia ovat esimerkiksi erilaiset aikasarjat säätilojen mittauspisteistä tai ainutkertaisen kulttuuriperinnön taltiointi digitoidussa muodossa.

Suomessa onkin parhaillaan meneillään opetus- ja kulttuuriministeriön Avoin tiede ja tutkimus -hanke, joka tähtää myös tärkeäksi tunnistettujen tutkimusdata-aineistojen pitkäaikaiseen säilyttämiseen. Kirjastoilla ja museoilla on käynnissä omien aineistojensa digitointi- ja tallennustyöt.

Datan arvon lisääntyneen ymmärryksen myötä on havahduttu pohtimaan kaupallisten julkaisijoiden tarkoitusperiä. Monet johtavat tieteelliset julkaisut vaativat nykyisin julkaisuun liittyvän datan itselleen, jolloin niille kerääntyy valtavasti maksumuurin taakse jäävää arvokasta tietoa.

Muun muassa suomalaishankkeet ajavat puolestaan *open data* -politiikkaa, jolla pyritään pitämään tieto saatavilla ja muidenkin tutkijoiden hyödynnettävinä.

”Datan hallinnasta ja tallennuksesta on yhä suurenevassa määrin tulossa osa tiedepolitiikkaa”, Lehtovuori tiivistää.

Kaksi kompastuskiveä

Kokeellisesti todistetusta käyttökelpoisuudesta huolimatta dna:n käyttöönoton esteenä on kaksi ongelmaa: nopeus ja hinta.

Piipohjaisiin, mikrosekuntiskaalassa toimiviin alustoihin verrattuna dna:n käyttäminen on tuskastuttavan hidasta. Suurin pulma on dna-syn-

Tara Brown / Washingtonin yliopisto

Pikkuruinen vaaleanpunainen läikkä ampullin pohjalla sisältää koodattuna sata maailmankirjallisuuden klassikkoteosta.

Kahdeksan teratavun kovalevyjä tarvitaan lähes 27 000, jotta päästään samaan tallennuskapasiteettiin kuin grammalla dna:ta.

teesin verkkaisuus. Käytännön sovelluksia varten synteesin tulisi olla liki satatuhatta kertaa nopeampi.

Historian perusteella voidaan laskea, että nopeutumisen kehitysvauhti on ollut pari kertalukua vuosikymmenessä. Näin on arvioitavissa, että tekniikka saavuttaa käyttökelpoisen tason parissa vuosikymmenessä.

Erlich ja Zielinski ilmoittavat tallennuskokeensa hinnaksi 9 000 dollaria. Suurimmaksi kulueräksi osoittautui dna-synteesi, jonka hinta oli 3 500 dollaria megatavulta. Vastaava summa datan lukemisesta oli 1 000 dollaria.

Vaikka menetelmän käyttö on siis nykyisellään hyvin kallista, tutkijapari luottaa tulevaisuuteen. Dna-synteesikemian edistysaskeleet tai vaikkapa *quick and dirty* -oligosynteesin kehittäminen voivat avata uusia mahdollisuuksia kustannustehokkaalle dna-tallennukselle.

Ihmiskunnan tuottaman digitaalisen datan on arvioitu saavuttavan vuoteen 2020 mennessä 44 tsettatavun (10^{21}) määrän. Tämä edellyttää mittavaa teknologiamurrosta datan säilytyksessä.

Dna:n valintaan datavarastoksi sisältyy sen muiden mainioiden ominaisuuksien lisäksi käänteentekevä periaatteellinen piirre.

Ihmisen luoma teknologia tuhoaa huolestuttavassa määrin tarpeellisia resursseja, elinympäristöä ja elämää. Yhä enemmän painoarvoa saa argumentti, jonka mukaan teknologiamme tulisi muuttua mahdollisimman paljon luontoa jäljitteleväksi.

Dna:n käyttäminen tietotekniikassa olisi erinomainen esimerkki tämän uuden teknologiaparadigman mukaisesta ajattelusta. ❐

Kirjoittaja on kemisti ja vapaa toimittaja.

Tietotulva ajaa kovalevyt ahtaalle

Toiminnassa on yhä data-arkistoja, jotka hyödyntävät vanhaa magneettinauhoihin perustuvaa teknologiaa. Vaikka nauhojen tallennuskapasiteetti on rajallinen, niillä on myös hyvä puolensa: ne eivät lepotilassa käytä energiaa.

Paljon yleisempiä ovat kiintolevyihin eli kovalevyihin perustuvat muistijärjestelmät, mutta ne kuluttavat energiaa lepotilassakin, tuottavat lämpöä ja vaativat siten konetilan jäähdytystä.

Kovalevyjen teknologiaa kehitetään jatkuvasti. Nykyään saatavilla olevien 3,5-tuumaisten levyjen kapasiteetit vaihtelevat kahden ja kahdeksan teratavun välillä, ja markkinoille tekevät tuloaan uuden teknologian kymmenen teratavun levyt.

Lähitulevaisuus lupaa vielä uudemmilla teknologioilla toimivia 12 ja 14 teratavun levyjä. Vaikka uusi teknologia on aina vanhaa kalliimpaa, lisääntynyt tallennuskapasiteetti kompensoi levyjen hintojen nousua.

Levyjä voidaan kytkeä toisiinsa kehikoissa, joihin niitä yleensä mahtuu muutamista kymmenistä liki sataan. Tällöin kehikon tilakapasiteetti voi olla useita petatavuja. Kehikoita voidaan edelleen kytkeä toisiinsa suuremmiksi systeemeiksi.

Tällaisten järjestelmien ostohinta ja ylläpitokustannukset ovat sitä luokkaa, että niihin on varaa vain suurilla datakeskuksilla. Jatkuvasti kasvavan lisätilan tarpeen ahdistamat toimijat joutuvat silti päivittämään järjestelmiään noin viiden vuoden välein.