

De sciencefiction voorbij: dataopslag in DNA

door Jasper Snoeren

De huidige generatie digitale informatiedragers zal langzaam maar zeker uit de gratie raken, want ze zijn niet duurzaam genoeg en hebben een beperkte opslagcapaciteit. Maar opslaan in DNA komt eraan. Hoewel de eerste toepassingen veelbelovend zijn, moeten er nog wel wat obstakels overwonnen worden.

Internetsnelheden nemen alsmaar toe, en daarmee ook de hoeveelheid data die wereldwijd wordt gecreëerd. Iedere kattenvideo, iedere vakantiefoto en ieder boodschappenlijstje wordt digitaal bewaard en gedeeld. Daardoor zijn we haast ongemerkt voor een van de grootste uitdagingen in ons vakgebied komen te staan. Want waar de behoefte aan dataopslag blijft toenemen, zal de opslagcapaciteit van digitale dragers uiteindelijk tegen een maximum aanlopen. Deze dragers hebben bovendien een beperkte levensduur: magnetische tape gaat slechts dertig jaar mee, een harde schijf kun je al na tien jaar vervangen. Kortom, we hebben een nieuw hulpmiddel nodig om data compacter en duurzamer op te slaan. Een oplossing hiervoor blijken we al sinds het begin der tijden bij ons te dragen: DNA.

Vele voordelen

DNA (desoxyribonucleïnezuur) bevat alle erfelijke informatie van een organisme. Het vormt de blauwdruk voor de eiwitten die alle processen in het lichaam regelen. DNA heeft de handige eigenschap dat het enorme hoeveelheden complexe informatie kan opslaan in een extreem klein formaat. Als opslagmedium voor data heeft het dus een gigantisch potentieel: theoretisch kan er circa 215 petabyte per gram DNA worden bewaard. Dat is pakweg het volledige archief van het Nederlands Instituut voor Beeld en Geluid in een rijstkorrel. Maar niet alleen het formaat biedt voordelen, het is ook buitengewoon duurzaam. Het DNA uit fossielen van mammoeten of uit het lichaam van ijsmummie Ötzi is nog steeds leesbaar. Wanneer het koel en droog bewaard wordt, gaat DNA honderdduizenden jaren mee zonder dataverlies. Bovendien is het energie-efficiënt: de moleculen zelf verbruiken geen elektriciteit. Ook zal de kennis over het gebruik van DNA niet verdwijnen zolang de genetische informatie van de mens erin is opgeslagen. De techniek om het materiaal te maken, bewerken en ontcijferen zal over honderd jaar ongetwijfeld verbeterd zijn, maar de informatie blijft onveranderd leesbaar.

Geen sciencefiction

Dat klinkt allemaal nogal vrij futuristisch. Toch is DNA als opslagmedium niet langer sciencefiction. In de afgelopen vijftien jaar zijn grote stappen gezet: als proof of concept zijn inmiddels al heel wat boeken, muziek en zelfs videoclips in DNA opgeslagen. Zo hebben onderzoekers van de Universiteit van Washington en Microsoft honderd boeken van Project Gutenberg omgezet in DNA, evenals de Universele Verklaring van de Rechten van de Mens in alle talen. En niet te vergeten de geluidsopnames op de gouden grammofoonplaat die in 1977 met de Voyager de ruimte in werd gezonden als boodschap aan buitenaardse levensvormen. Op dit moment zijn de onderzoekers bezig om de gehele Engelstalige Wikipedia-database vast te leggen.

Hoe werkt het?

Hoe gaat opslag in DNA-vorm precies in zijn werk? Een DNA-streng is een polymeer, opgebouwd uit een opeenvolging van vier *nucleotiden*. Die worden aangeduid met de letters A, T, C en G (oftewel de organische moleculen Adenine, Thymine, Cytosine en Guanine). Het werkt in principe op dezelfde manier als de enen en nullen in een digitale bitstream. Kunnen we informatie digitaliseren, dan kunnen we het dus ook vertalen naar een nucleotiden-reeks. Vervolgens wordt die genetische code in een laboratorium gesynthetiseerd – de moleculen worden opgebouwd. Hierbij gaat het om kunstmatig gecreëerd materiaal, er komt geen levend organisme aan te pas. De strengen worden opgeslagen in een vloeibare oplossing en kunnen vervolgens relatief gemakkelijk worden bewaard en vervoerd. Het maken van kopieën is snel en bovenal goedkoop: onder de juiste chemische omstandigheden kopieert DNA zichzelf, net als in het menselijk lichaam gebeurt.

Data opsplitsen

In de praktijk is het vertalen van een digitale code naar de juiste nucleotidenreeks echter iets ingewikkelder dan het op het eerste gezicht lijkt. Een DNA-streng mag bijvoorbeeld niet te lang zijn, en herhaling van te vaak dezelfde nucleotide achter elkaar vormt een risico op breuken. Om dit te voorkomen wordt één stuk data opgesplitst in tal van kleinere sliertjes DNA.

Ieder sliertje wordt voorzien van een zogenaamde primer. Die primer geldt als unieke identifier voor dat stuk data, en bevat ook aanvullende metadata over bijvoorbeeld de volgorde waarin de sliertjes gezet moeten worden om de data uit te lezen. Ieder sliertje komt honderden keren voor in het DNA-mengsel en bevat bovendien veel overlap met het voorgaande en opvolgende stukje bitstream. Op die manier draagt genetische data eigen back-ups mee en wordt de kans op lees- en kopieerfouten drastisch verminderd.

Hindernissen

De techniek om data in DNA op te slaan gaat met sprongen vooruit. Maar om het ook echt in de praktijk te kunnen gebruiken, moeten nog verschillende obstakels overwonnen worden. Het is hoofdzakelijk nog een heel langzaam proces. Het duurt nu vele uren om een bestand te openen (anders gezegd: het ontcijferen en digitaal weergeven van de nucleotidenreeks).

Dit weegt niet op tegen de microseconden die een flash drive voor dezelfde taak nodig heeft. De verwachting is dat er in de toekomst snelheden tot wel 100 Gbps behaald gaan worden. Tot die tijd is DNA-opslag alleen geschikt als een permanente, extreem duurzame back-up voor data die niet voortdurend geraadpleegd hoeft te worden. Denk bijvoorbeeld aan de kennis die niet verloren mag gaan bij een grote klimaatramp of wereldoorlog. Maar de belangrijkste drempel is het prijskaartje. De kosten voor het decoderen en synthetiseren van DNA mogen dan al rap dalen, het is nog steeds een prijzige aangelegenheid. De bedragen voor het synthetiseren van enkele MB's aan data lopen al gauw op tot een paar duizend dollar. Daar durft een archief in de culturele sector nog niet eens van te dromen.

Zelfs wanneer het schrijven van DNA betaalbaar wordt, zullen de serverkasten nog niet direct aan de straat gezet worden. Transcoderen van digitale naar genetische data is één ding, maar het vervangen van de huidige infrastructuur in de datacenters is eveneens een immense klus. Commerciële leveranciers zullen er niet zomaar aan beginnen zodra de techniek ook buiten de wetenschappelijke wereld voorhanden komt.

Kwestie van tijd

Volgens schattingen gepubliceerd door *Nature* zal het wereldwijde digitale archief in 2040 zo extreem veel data bevatten, dat de opslag ervan op huidige flash memorysticks honderd keer meer silicium van microchip-kwaliteit nodig heeft dan er naar verwachting op aarde voorradig is. Dat DNA data-opslag voor de oplossing gaat zorgen lijkt gelukkig slechts een kwestie van tijd. De onderzoekers van Microsoft verwachten dat het

hooguit een decennium zal duren voordat de techniek echt klaar is om de wereld te veroveren. Wordt de informatieprofessional straks ook bio-chemicus?

- Het wereldwijde digitale archief in 2040 zal zo extreem veel data bevatten, dat de opslag ervan honderd keer meer silicium van microchip-kwaliteit nodig heeft dan er op aarde voorradig is. Andere hulpmiddelen zijn dus hard nodig. Een veelbelovende oplossing dragen we van nature al bij ons: DNA.
- In DNA kan data extreem compact worden opgeslagen. Wanneer het koel en droog bewaard wordt, gaat DNA bovendien honderdduizenden jaren mee zonder dataverlies. Ook is het erg energie-efficiënt: de moleculen zelf verbruiken geen elektriciteit.
- Er zijn nog wat praktische obstakels te overwinnen, maar het zal geen tien jaar meer duren voordat data-opslag in DNA op de markt beschikbaar komt.

Naschrift

Als eerste kennismaking met de technologie heeft Beeld en Geluid, in samenwerking met Twist Bioscience, recentelijk de iconische beelden rondom de viering van het gewonnen EK '88 in DNA gearchiveerd (zie <http://pers.beeldengeluid.nl/192662-beeld-en-geluid-slaat-iconisch-fragment-ek-88-op-in-dna>).

Bronnen

- Bornholt, J., Lopez, R., Carmean, D., Ceze, L., Seelig, G. & Strauss, K. (2016). ASPLOS 2016, International Conference on Architectural Support for Programming Languages and Operating Systems. Gepubliceerd door ACM, *Association for Computing Machinery*.
- Extance, A. (2016). How DNA could store all the world's data. *Nature*. Geraadpleegd op <https://www.nature.com/news/how-dna-could-store-all-the-world-s-data-1.20496>.
- Kolata, G. (2017). Who Needs Hard Drives? Scientists Store Film Clip in DNA. *New York Times*. Geraadpleegd op <https://www.nytimes.com/2017/07/12/science/film-clip-stored-in-dna.html>.
- Milenkovic, O., Gabrys, R., Kiah, H.M. & Hossein Tabatabaei Yazdi, S.M. (2018). Exabytes in a Test Tube: The Case for DNA Data Storage. *IEEE Spectrum*, IEEE. Geraadpleegd op <https://spectrum.ieee.org/semiconductors/devices/exabytes-in-a-test-tube-the-case-for-dna-data-storage>.
- Monaghan, P. (2019). A solution for all digital storage woes? *Moving Image Archives News*. Geraadpleegd op <http://www.movingimagearchivenews.org/a-solution-for-all-data-storage-woes/>.
- Rosenthal, R. (2018). DNA's niche in the storage market. Geraadpleegd op <https://blog.dshr.org/2018/02/dnas-niche-in-storage-market.html>.
- Yong, E. (2017). Scientists Can Use CRISPR to Store Images and Movies in Bacteria. *The Atlantic*. Geraadpleegd op <https://www.theatlantic.com/science/archive/2017/07/scientists-can-use-crispr-to-store-images-and-movies-in-bacteria/533400/>.

Eerder gepubliceerd in: *IP – vakblad voor informatieprofessionals*, nr. 7/2019. Zie ook: https://informatieprofessional.nl/resources/uploads/2019/10/IP2019_07-reDeel.pdf