



Kopje onder in de data

Het sequensen van een genoom of exoom wordt met de dag goedkoper, maar de **dataopslag** is kostbaar en software om slimme analyses te doen is schaars. Wat te doen met al die data?

KEVIN KOSTERMAN

Zo'n 10 jaar geleden is eindelijk het eerste menselijke genoom geheel in kaart gebracht. Het duurde 13 jaar om de ongeveer 3 miljard basenparen op een rijtje te zetten. Met de huidige technieken en capaciteiten kan dit in een dag. Daarom sequensen steeds meer onderzoeksgroepen exomen en genomen, op zoek naar genetische defecten die verantwoordelijk zijn voor ziekten en naar bepaalde eigenschappen van soorten.

Die ontwikkeling gaat gepaard met steeds meer data: ongeveer 1,5 terabyte per genoom. Het DNA wordt namelijk in stukjes van 50 tot 150 basenparen geknipt; voor een betrouwbare weergave van het genoom moet elk van die stukjes zo'n 40 tot 80 keer door de analysemolien heen.

Peter van der Spek, professor bioinformatica van het Erasmus MC: "Er komen nog heel veel onbekende varianten voor in het genoom. Om er zeker van te zijn dat zo'n variant geen leesfout is van het appa-

raat, bekijken we niet alleen die tientallen runs maar vergelijken we ze ook met het genoom van honderden andere mensen."

Ook bijvoorbeeld genetisch kankeronderzoek is vergelijkend van aard: je vergelijkt tumor-DNA met normaal DNA, of je zoekt naar de mate van erfelijkheid

'Opnieuw sequensen is veel goedkoper dan alles opslaan'

door het genoom van meerdere mensen naast elkaar te leggen. Iedere vergelijking is goed voor diezelfde 1,5 terabyte aan informatie.

CRASHENDE COMPUTERS

Joris Vermeesch, professor genomics van de K.U.Leuven, ervaart soms de problemen van de hoeveelheid data. "Af en toe crashen onze computers en dat komt

puur door capaciteitsgebrek. Ondanks dat we een supercomputer hebben. Het probleem is dus de opslag." Volgens zijn collega Stein Aerts, hoofd van het laboratorium voor Computationale Biologie, vallen de problemen tegenwoordig echter mee. "Het is zeker een uitdaging om met de grote hoeveelheden data om te gaan, maar er is steeds meer gestandaardiseerd. De informatie uit de sequenser is niet meer zo exotisch als enkele jaren geleden. De grootste uitdaging is nu het identificeren van de mutaties die ten grondslag liggen aan bepaalde ziekten. Maar inderdaad is daarvoor vaak data-integratie nodig, en die analyses kunnen maanden in beslag nemen."

Datareductie levert mogelijk een voordeel op. Is het sequensen van alleen de exomen dan de oplossing? Dit zou een besparing van bijna 99 procent kunnen opleveren. Van der Spek: "Alleen het sequensen of opslaan van de exomen heeft naar mijn mening niet veel nut. Je mist dan bijvoorbeeld de bindingsplaatsen



Niet iedereen heeft een supercomputer tot zijn beschikking.

van transcriptiefactoren in de niet-codeerende gebieden. Die hebben een subtiele invloed op veel biologische processen. Daarnaast blijken de ultrageconserveerde elementen die je dan mist, een belangrijke rol te spelen bij celdeling – zeer relevant dus bij kankeronderzoek.”

REFERENTIEGENOOM

Moeten we dan altijd een volledig genoom opslaan? Nee. Een groot deel van die ruwe informatie is namelijk voor alle mensen gelijk. Zelfs een chimpansee komt genetisch gezien voor 96 procent overeen met een mens. Met referentiegenomen heb je het merendeel van de informatie dus al in huis. Van der Spek: “Pluk je twee willekeurige mensen van de Meir en vergelijk je hun genoom, dan vind je zo’n 2,3 miljoen SNP’s (single nucleotide polymorphisms, red.): verschillen ten opzichte van de referentie. Dat is de enige informatie die je hoeft op te slaan. Dat is een stuk minder dan de 6 miljard oorspronkelijke basenparen, namelijk 3 miljard basenparen met elk twee allelen. Je moet dan rekenen op zo’n 1,5 gigabyte aan data voor een volledig genoom ten opzichte van de referentie.”

Aerts: “Je slaat dus maar een fractie van de oorspronkelijke informatie op. Maar

we bewaren wel altijd de DNA-stalen, mochten die ooit nog nodig zijn. Dan is opnieuw sequensen namelijk veel goedkoper dan alles opslaan. Een terabyte aan opslagruimte betekent toch een kostenpost van circa 300 euro per jaar.”

Toch heb je dan nog vaak te maken met een onhandelbare hoeveelheid data. Abhishek Singh, onderzoeker van BBMRI (Biobanking and Biomolecular Resources Research Infrastructure) en betrokken bij

‘Een terabyte aan opslagruimte kost circa 300 euro per jaar’

het ‘genoom van Nederland’-project: “Ondanks dat je de data kunt comprimeren en referenties kunt gebruiken, houden we nog enorme hoeveelheden data over. Gecomprimeerd komen onze scans uit op zo’n 3 gigabyte per persoon; aan ons onderzoek doen nu 760 personen mee. Zonder een high performance-computer zoals we die in Groningen hebben, kun je daar nooit mee werken.”

Om de rekentijd terug te brengen schafte het Erasmus MC dit jaar een zogeheten

DATAGROOTTE:

- 1 terabyte (TB) = 1.000.000.000.000 bytes
- 1 gigabyte (GB) = 1000.000.000 bytes
- 1 kilobyte (kB) = 1000 bytes

Een A4tje volgetikt: 5 kB en 20 cm breed

Het gesequenste menselijke genoom (1,5 TB): 60.000 km aan A4tjes

Het menselijk genoom uitgeschreven (6 GB): 240 km aan A4tjes

Het menselijk genoom als uitgerold DNA-molecuul: 3 meter

De menselijke celkern met daarin het opgevouwen DNA: 2,5 micrometer



Oracle exadatamachine aan. Van der Spek: “Er zijn wereldwijd vier partnerships in de VS en een in Europa. Met deze computer kunnen we binnen enkele seconden over meer dan honderd genomen een zoekopdracht uitvoeren die oorspronkelijk een minuut of zes à zeven duurde. Nu zijn zes minuten niet zo lang, maar als je twintig posities in het DNA wil checken bij honderd mensen duurt dat dus uren.” Aerts: “Onze computer heeft nu een kracht van 100 CPU’s; een gemiddelde laptop heeft 2 CPU’s. Wij kunnen nu berekeningen uitvoeren waar je normaal niet op zou willen wachten”.

CLOUDS

Wie echter af en toe sequencingdata wil analyseren, beschikt natuurlijk niet over een exadatamachine of computer met 100 CPU’s. Daarom komen er tegenwoordig steeds meer zogeheten clouddiensten op de markt. Bij *cloud computing* vindt de opslag online plaats en is ook de analyse, en daarmee de benodigde rekenkracht, elders ondergebracht. Dat kan bijvoorbeeld bij Amazon Cloud en sinds kort ook bij BGI (het voormalige Beijing Genome Institute). Maar ondanks de snelle internetverbindingen van tegenwoordig kan het versturen van data nog steeds een probleem vormen. Aerts: “Ik zie momenteel nog niet de voordelen van cloud computing. Je bent dagen bezig met het uploaden en downloaden van de data.” Singh: “Snelheid kan een probleem zijn bij het up- en downloaden, maar voor gebruikers die af en toe iets willen doen kan dit een uitkomst bieden.”

Ook clouddiensten zijn dus nog niet zaligmakend; het blijft voorlopig een kwestie van wachten op nieuwe en betaalbare oplossingen.



Het menselijk genoom: 240 km aan volgeschreven A4tjes. Dit is hemelsbreed de afstand tussen Oostende en Bastenaken.