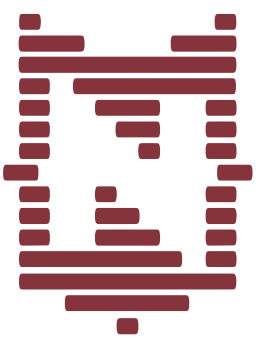
DNA dilemma

De puzzelstukjes zijn vijf verknipte kopieën
van eenzelfde DNA-streng.

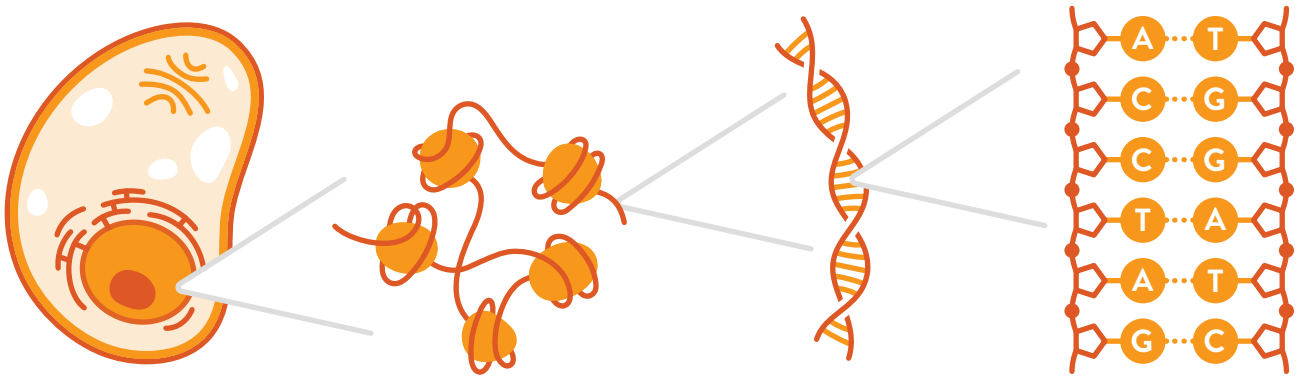
UITDAGING

Kun jij de oorspronkelijke
DNA-code reconstrueren?



DNA: de instructies die ons (min of meer) maken tot wie we zijn

De vormgeving van de puzzelstukken doet hopelijk een belletje rinkelen: de letters A, C, G, T doen al snel aan DNA denken. Inderdaad, deze puzzel is geïnspireerd op een praktisch probleem waar wetenschappers die met DNA werken geregeld tegenaan lopen.



De cellen in je lichaam bevatten instructies voor de taken die ze kunnen en moeten uitvoeren. Die instructies zijn geschreven in een bijzondere "taal" bepaald door de volgorde van vier bouwstenen in grotere DNA-moleculen: *adenine*, *cytosine*, *guanine* en *thymine*. De DNA-moleculen, met hun karakteristieke helixvorm, worden opgerold tussen heel wat steungevende eiwitten in de celkern bewaard, en ken je wellicht beter als onze *chromosomen*.

De code in je DNA zorgt er onrechtstreeks voor dat bijvoorbeeld de juiste eiwitten gemaakt worden om je spieren structuur geven, dat cellen enzymen maken om je eten te verteren, dat er celdeling optreedt zodat je lichaam zichzelf kan herstellen bij een kwetsuur, dat pigmenten in een bepaalde hoeveelheid gefabriceerd worden (die dan weer je haarkleur bepaalt), ...

Hoe die eiwitsynthese precies tot stand komt door DNA uit te lezen, valt helaas niet in een-twee-drie uit te leggen. Daarvoor verwijzen we naar deze uitstekende video.

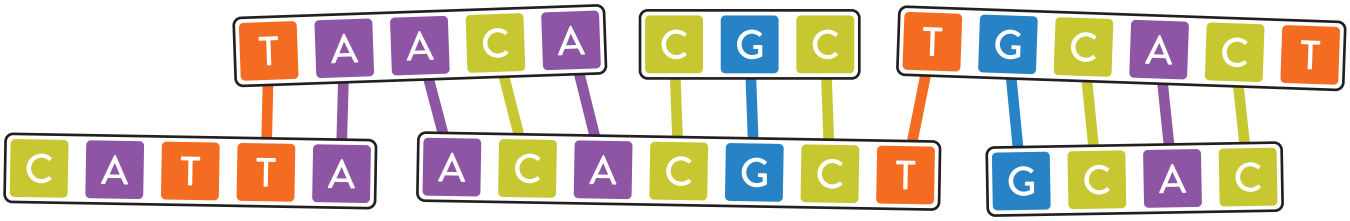
DNA uitlezen

Omwille van het belang van die code willen onderzoekers, dokters, maar bijvoorbeeld ook plantveredelaars die graag inkijken. Die deelmoleculen of basen vertalen naar een mooi van rijtje A's, C's, G's en T's noemt men *sequencing*.

Ondanks de veelheid aan complexe machines specifiek ontworpen voor die taak is sequencing absoluut geen evidente opgave. Omwille van biochemische redenen blijkt de betrouwbaarheid van het uitgelezen resultaat na enkele tien- tot honderdtallen basen (afhankelijk van methode) erg te gaan dalen, in die mate dat je evengoed kan gokken.

Zelfs state-of-the-art sequencingtechnieken raken niet verder dan een paar miljoen basen. Dat lijkt veel, tot je beseft dat de langste menselijke chromosomen bijna 250 miljoen basen lang zijn! Bovendien komen die moderne technieken met heel wat trade-offs, zoals een hoog prijskaartje en een veel grotere foutenmarge. De vraag stelt zich dus, *hoe kunnen we uit een hoop korte stukjes het volledige DNA reconstrueren?* Deze stap wordt ook wel *assembly* genoemd.

Uiteraard kun je met één verknipte kopie niks aanvangen; hoe zou je immers weten welk stuk na welk ander stuk komt? Daarom werkt men met meerdere kopieën, die min of meer willekeurig verknipt worden. Aan de hand van de overlap tracht men dan de volgorde te reconstrueren.



Overlap zoeken en bij elkaar plakken, klinkt niet zo heel moeilijk, maar je merkt ongetwijfeld dat de puzzel toch niet evident is. Bedenk je dan dat je voor het uitlezen van één menselijk chromosoom geen 35 stukken verdeeld over vijf kopieën moet samenpuzzelen … maar eerder zo'n 300 miljoen stukjes over 30 kopieën!

Daar bestaan toch computers voor?

Gezien de complexiteit van de vraag én de ronduit absurde hoeveelheid data, zou het (exact) reconstrueren van zo'n chromosoom om en bij de $10^{15\,000\,000}$ jaar duren — dat is een getal met 15 miljoen cijfers. Laat het duidelijk zijn: ook met computers is dit geen optie!

Er zijn goede redenen om te geloven dat er simpelweg geen efficiënter algoritme is voor DNA-assembly: het probleem is immers NP-compleet. Een mooie introductie vind je hier.

DNA assembly in de praktijk

Desondanks hoor je geregeld over geneticastudies, het *human genome project*, … Hoe doet men dat dan als het zo onredelijk moeilijk en tijdsintensief is? Voornamelijk gaat men aan de slag met benaderende algoritmes: men is in feite tevreden met een "goede gok" in plaats van een echt exacte assembly te zoeken.

Ook dan blijft het een rekenintensief proces en gaat men dus allerlei andere truken gebruiken. Zo wordt er nog altijd geoptimaliseerd aan de bestaande programma's, gebruikt men supercomputers, probeert men te vergelijken met een gekend gelijkaardig genoom, beperkt men zich tot een kleiner stuk DNA, … of leest men het DNA gewoon helemaal niet uit en gebruikt men zogenaamde SNP-arrays.

Alsof het probleem nog niet moeilijk genoeg was, moet men ook nog eens rekening houden met de slordige realiteit van DNA. Zo is sequencing zeker niet foutloos, ook niet bij korte stukjes. Bovendien zijn de begin- en eindstukken vaak net het meest foutgevoelig. Dat wil ook zeggen dat referentiegenomen hoogstwaarschijnlijk niet zonder fouten zijn. Daarenboven zijn de meest frequente variaties binnen eenzelfde diersoort de moeilijkste om van fouten te onderscheiden: ofwel zijn ze erg klein (vaak slecht één aangepaste base), ofwel erg onopvallend (zoals zeven in plaats van zes repetities van eenzelfde stukje).

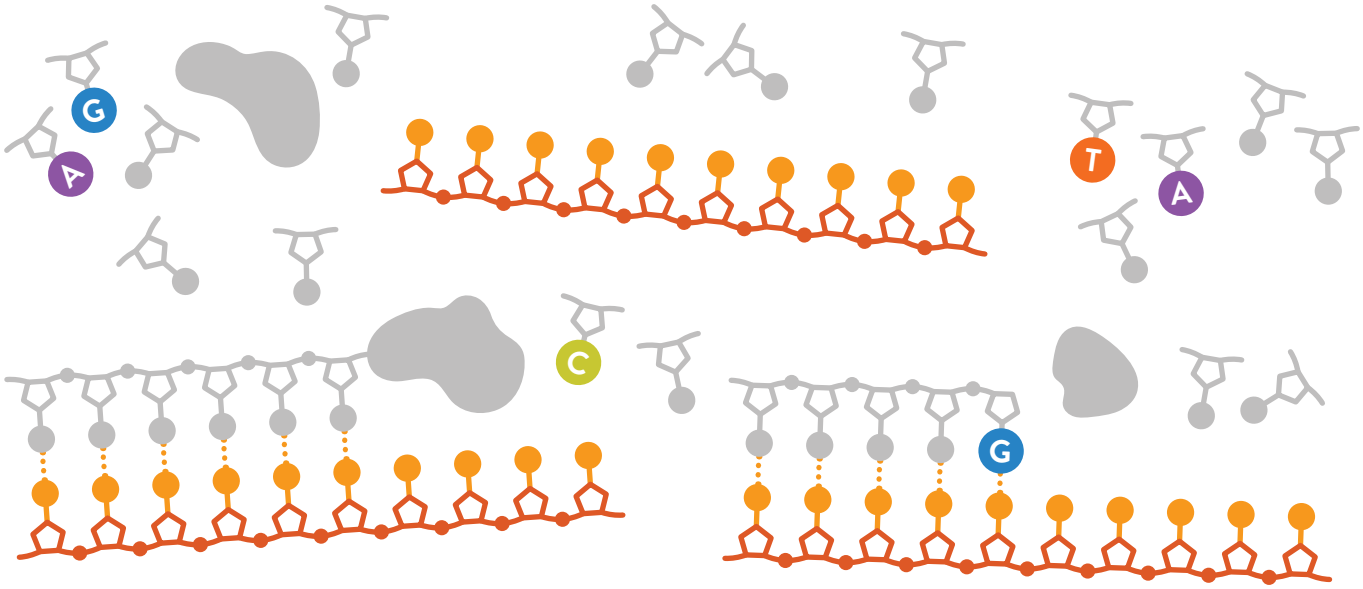
Hoe komen we überhaupt aan die kortere stukjes A, C, T, G?

Er zijn heel veel versies en varianten van het proces om uit een biologisch monster digitale data te halen. Veel van de details zijn bedrijfsgeheimen en/of ongelooflijk technisch, maar voor de geïnteresseerde lezer beschrijven we het conceptuele plaatje. Hetzelfde kopiëren, fluoresceren en illumineren ligt namelijk nog altijd aan de basis van nieuwe volautomatische toestellen.

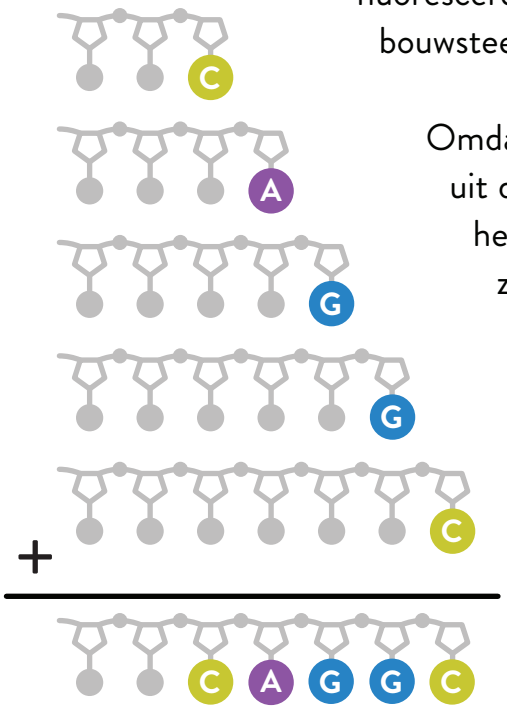
Men start met het isoleren van DNA uit een monster, het op te delen in geschikte kleine stukjes en elk stukje voldoende keren te kopiëren. Een van de meest courante technieken daarvoor is PCR … een term waar je recent moeilijk aan kon ontsnappen!

Hoe PCR juist werkt en hoe de techniek gebruikt wordt in de beruchte PCR-testen, kun je hier bekijken.

Die hoop kopieën wordt nog een laatste keer uitgesplitst tot enkelvoudige strengen en gekopieerd, maar onder speciale omstandigheden. Men maakt een biologisch soepje met het DNA, de nodige enzymen ("werkertjes"), voldoende bouwstenen A, C, G en T, én enkele speciale bouwstenen met een fluorescerend "staartje". Door die speciale A, C, G en T elk met hun eigen kleur te fluoresceren, kan men die later op macroscopische schaal nog herkennen.



De enzymen in dat soepje zullen nieuwe kopieën beginnen aanmaken, maar telkens ze toevallig zo een gefluoresceerde bouwsteen gebruiken, loopt dat kopiëren vast. Uiteindelijk vindt men zo een boel segmentjes van variabele lengtes: de ene is wat verder gekopieerd dan de andere. Door naar het fluorescerende kleur te kijken, kan men bovendien die laatst gebruikte bouwsteen aflezen.



Omdat elk enzym aan hetzelfde startpunt vertrokken is, kan men uit de kleur en lengte van ieder kopietje één letter en positie in het oorspronkelijke DNA-segment afleiden. Een segment van zes basen lang dat groen fluoresceert, kan bijvoorbeeld aanduiden dat cytosine op positie zes voorkomt. Uiteraard zijn DNA-segmenten te klein om met de lat op te meten, dus in de plaats gaat men ze sorteren van klein naar groot (via gelelektroforese). Zo kan men de DNA-sequentie in feite gewoon aflezen.

Meer weten?

Zoals deze puzzel illustreert, kost DNA lezen heel wat moeite en bijgevolg veel geld en tijd. Als jij een genetische test krijgt, gaat men daarom bijna altijd met een SNP-array werken, iets wat commerciële DNA-kits maar al te graag onder de mat vegen. Zo'n SNP-array bevat in ieder vakje een specifiek stukje DNA-streng dat met één versie van een gekende SNP-locatie overeenkomt. Door dit gecontroleerd te laten reageren met een kopie van je eigenlijke DNA, kan je dan uitlezen of jij die specifieke versie of net de andere hebt, zonder het vele werk om heel je DNA uit te lezen. Maar je moet natuurlijk wel veel concreter weten waarnaar je zoekt opdat deze techniek nuttig zou zijn!

- CD Genomics, *The principles and workflow of SNP microarray.*

Het *Human Genome Project* doet hoogstwaarschijnlijk een belletje rinkelen: een ambitieus internationaal project opgestart in 1990, met als doel om binnen de 15 jaar het volledige menselijke genoom in kaart te brengen. In 2003 werd het afgerond verklaard … maar eerder dit jaar kwam het nieuws dat het menselijk genoom nu pas "eindelijk compleet" is. Sommige stukken DNA bleken zoveel moeilijker dan andere om te sequencen (en wellicht niet zo belangrijk), dat men op een gegeven moment besloot om te stoppen.

- Viktor McElheny, *Drawing the map of life: inside the Human Genome Project.* Basic Books, 2012.
- De Morgen, 31 maart 2022, *Genoom van de mens is eindelijk compleet (en dat brengt meteen verrassingen aan het licht).*

Onze puzzle demonstreert eigenlijk wat er gebeurt bij *de novo single end sequencing*, maar je kan bijvoorbeeld ook werken met *paired end reads*, referentiegenomen … Een mooie introductie van al die verschillende smaken aan DNA sequencing vind je hier.

- Your Genome, *How do you put a genome back together after sequencing?*

PCR is trouwens niet de enige optie om DNA te kopiëren. Ook andere technieken zoals BAC's waarbij men gebruik maakt van bacteriën, kunnen in bepaalde situaties nuttig zijn.

- US National Human Genome Reserach Institute, *Bacterial artificial chromosome (BAC).*

Hoewel dit vakgebied niet zonder biologen én computerwetenschappers kan, vliegen ze elkaar soms wel eens in de haren. Om te begrijpen waarom kunnen we deze blogpost wel aanbevelen. Er wordt uitgelegd waarom theoretische papers over NP-compleetheid van DNA-problemen soms nogal in het verkeerde keelgat schieten bij biologen, waarom ze vaak concreet maar weinig bijdragen voor praktische problemen, en waarom ze tóch relevant zijn.

- Bits of Bioinformatics, *Does NP-completeness have a role to play in Bioinformatics?*