Bioinformatica: een nieuw vakgebied dat drijft op toegepaste wiskunde en statistiek

Geïnspireerd door het verhaal van Miranda van Uitert (zie pag. 2 en 3) formuleerde Sven Warris, medewerker aan het Bioinformatics Expertise Center van de Hanze University Groningen, de volgende opdracht voor leerlingen van 5-havo en 5/6-vwo.

Elke cel van de mens bevat DNA: het kookboek dat beschrijft hoe die cel moet functioneren. Als DNA het kookboek is, dan is het gen het recept. Is een recept correct, dan werkt de cel goed, maar als de beschrijving foutjes bevat, gaat het verkeerd: de cel functioneert niet goed. Dat kan leiden tot ziekten zoals kanker.

De afgelopen jaren zijn de technologische mogelijkheden voor DNA-onderzoek enorm toegenomen. Steeds meer gegevens komen steeds sneller beschikbaar. Deze grote hoeveelheid gegevens is door onderzoekers of artsen niet meer te overzien. Computers nemen de analyse van de gegevens over.
Het vakgebied dat hierdoor is ontstaan, heet bioinformatica. Wiskunde, statistiek en informatica zijn belangrijke hulpmiddelen geworden bij het onderzoek naar ziekten en bij de ontwikkeling van geneesmiddelen en gezondheid. Meer informatie hierover vind je op http://www.watisgenomics.nl.

Deze volgende opgaven geven een indruk van de rol van wiskunde en statistiek in het bioinformaticaonderzoek.

Opdracht 1

DNA bestaat uit chemische bouwstenen, zogenaamde nucleotiden of basen. Zoek op uit hoeveel nucleotiden het menselijk DNA bestaat. Vergelijk dit met het aantal nucleotiden van het DNA van een varken. Hoeveel groter is het menselijk DNA? Laat dit zien dat een mens dus geen varken is?

Opdracht 2

Tien nucleotiden achter elkaar hebben een lengte van 10 Å (Ångstrom). Wat is de lengte van het DNA van een menselijke cel? Zoek op uit hoeveel cellen je lichaam bestaat. Wat is naar schatting de totale lengte van al het DNA uit je lichaamscellen? Hoeveel keer de afstand tot de zon en terug is dat?

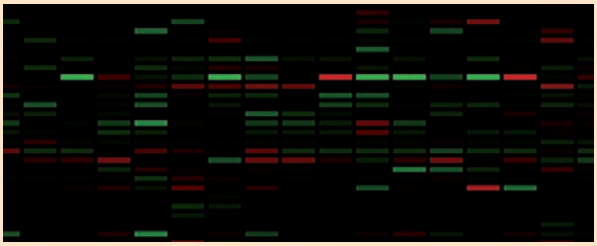
Cellen zijn geprogrammeerd om eiwit te maken. Dit gebeurt buiten de celkern. De informatie uit de celkern die hierbij nodig is, wordt overgebracht door RNA. De biologische informatieverwerking in een cel loopt dus via de reeks

DNA (kookboek) ⟹ RNA (recept) ⟹ eiwit (gerecht)

Opdracht 3

Zoek uit hoeveel menselijke genen (dus stukjes DNA die de aanmaak van eiwitten regelen) op dit moment bekend zijn.
Een veelgebruikte techniek om de werking van DNA te bestuderen en ziek en gezond te vergelijken is de *microarray*. Dit is een glasplaatje ('chip') met daarop stukjes DNA. Bij onderzoek met microarrays wordt RNA uit zowel ziek als gezond weefsel geïsoleerd. Vervolgens wordt er gemeten hoeveel RNA er van ieder gen in de twee toestanden (ziek/gezond) aanwezig is.

De hoeveelheid RNA van een gen noemen we de *expressie* van het gen. Met microarrays bepaal je dus de expressie van genen. Veranderingen in de expressie - en dus de activiteit - van een gen, kunnen een aanwijzing zijn voor de betrokkenheid van dat gen bij de totstandkoming van te onderzoeken ziekten. Om te bepalen of een gen meer dan wel minder actief geworden is, bereken je de *expressieratio*. Dit is de verhouding tussen de expressie in de zieke cel ten opzichte van de gezonde cel.



Detail van een microarray. De kleuren geven aan hoe actief genen zijn ten opzichte van een referentie-niveau. Groen: sterk gedeactiveerde genen, rood: sterk geactiveerde genen.

In dit onderzoek bekijken we borstkankercellen. We willen weten waarom en wanneer een borstcel zich ontwikkelt tot een kankercel. Daarvoor doen we zeven experimenten waarin we borstcellen in verschillende stadia van de ontwikkeling van een tumor onderzoeken.

Opdracht 4

Elk gen komt twee keer voor op een microarray. De activiteit van elk gen meet je met behulp van RNA. Hoeveel gegevens krijg je na zeven experimenten waarin je steeds gezond en ziek weefsel - dus op twee verschillende arrays - vergelijkt?

Opdracht 5

Het kost ongeveer een seconde om met de rekenmachine de expressieratio van een gen uit te rekenen. Hoeveel werkdagen heb je nodig om op die manier alle ratio's te berekenen?

De expressieratio's leveren een getal op tussen -1 en 1.

-1 = een gen is sterk gedeactiveerd in de zieke cel
0 = er is geen verschil tussen de zieke en de gezonde cel
1 = een gen is sterk geactiveerd in de zieke cel

Genen die in al onze experimenten hetzelfde verschil geven tussen ziek en gezond (bijvoorbeeld steeds geactiveerd) zouden heel goed een rol kunnen spelen bij het ontstaan van borstkanker. We willen dus bepalen welke genen qua expressiepatroon op elkaar lijken en welke niet.
Dit 'op elkaar lijken' is een wiskundige maat die we de afstand noemen. Is de afstand nul, dan vertonen genen hetzelfde expressiepatroon over alle experimenten. Hoe groter de afstand, hoe verder de patronen uit elkaar liggen. Bij dit soort onderzoek wordt veel gebruik gemaakt van de Euclidische afstand. Voor punten $P(p_x, p_y)$ en $Q(q_x, q_y)$ in een tweedimensionale ruimte is deze gedefinieerd als:

$$d = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}$$

Opdracht 6

Uit hoeveel dimensies bestaat een punt in dit onderzoek?

Opdracht 7

Geef de formule om de Euclidische afstand tussen de expressie van twee genen te berekenen in een dergelijke x-dimensionale ruimte (x = het antwoord op opdracht 6).

Opdracht 8

Hoeveel afstanden moeten we berekenen om de hele dataset te kunnen analyseren?

Opdracht 9

Een computer berekent binnen 0.0001 seconde de afstand tussen twee genen. Hoe lang heeft een computer nodig om alle afstanden te berekenen?

Met behulp van de complete afstandsmatrix kunnen we nu groepen maken van genen met eenzelfde expressiepatroon. We doen dit met een wiskundig proces dat *clusteren* heet. Hiervoor bestaan verschillende methoden. Wij gebruiken de hiërarchische clustermethode. Die werkt als volgt. Begin met twee genen die het dichtst bij elkaar liggen en combineer die in een cluster. Neem daarna een volgend gen. Als je wilt weten of dit gen dichtbij een cluster van genen ligt, neem je het gen uit het cluster dat er het dichtst bij ligt. Herhaal dit steeds opnieuw. Deze procedure heet *single linkage clustering*. Kijk voor een voorbeeld op http://en.wikipedia.org/wiki/Data_clustering.

Opdracht 10

Neem de volgende fictieve expressiedata.

Gennaam	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	Exp. 6	Exp. 7
ACTA1	0,30	0,26	-0,11	0.28	0,30	0,21	0,29
PGLYRP3	-0,90	-0,40	0,13	0,02	-0,57	-0,84	-0,66
BRCA2	0,68	0,59	0,74	0,81	0,94	0,77	0,84
BRMS1	0,59	0,64	0,79	0,62	0,85	0,67	0,79

De tabel geeft de expressieratio's van vier genen in zeven opeenvolgende stadia van een borsttumor. Bereken de Euclidische afstand tussen elk tweetal genen en gebruik deze om de genen te clusteren. Wat valt je op? Wat zegt dat over de mogelijke betrokkenheid van deze genen bij borstkanker?

Opdracht 11

Zoek op internet de betekenis en functie van de verschillende gennamen op. Wat valt je op?