Klaas-Jan Slooten
Nederlands Forensisch Instituut
Postbus 24044
2490 AA Den Haag
k.slooten@nfi.minvenj.nl

Onderzoek

Kansrekening in forensisch DNA-onderzoek

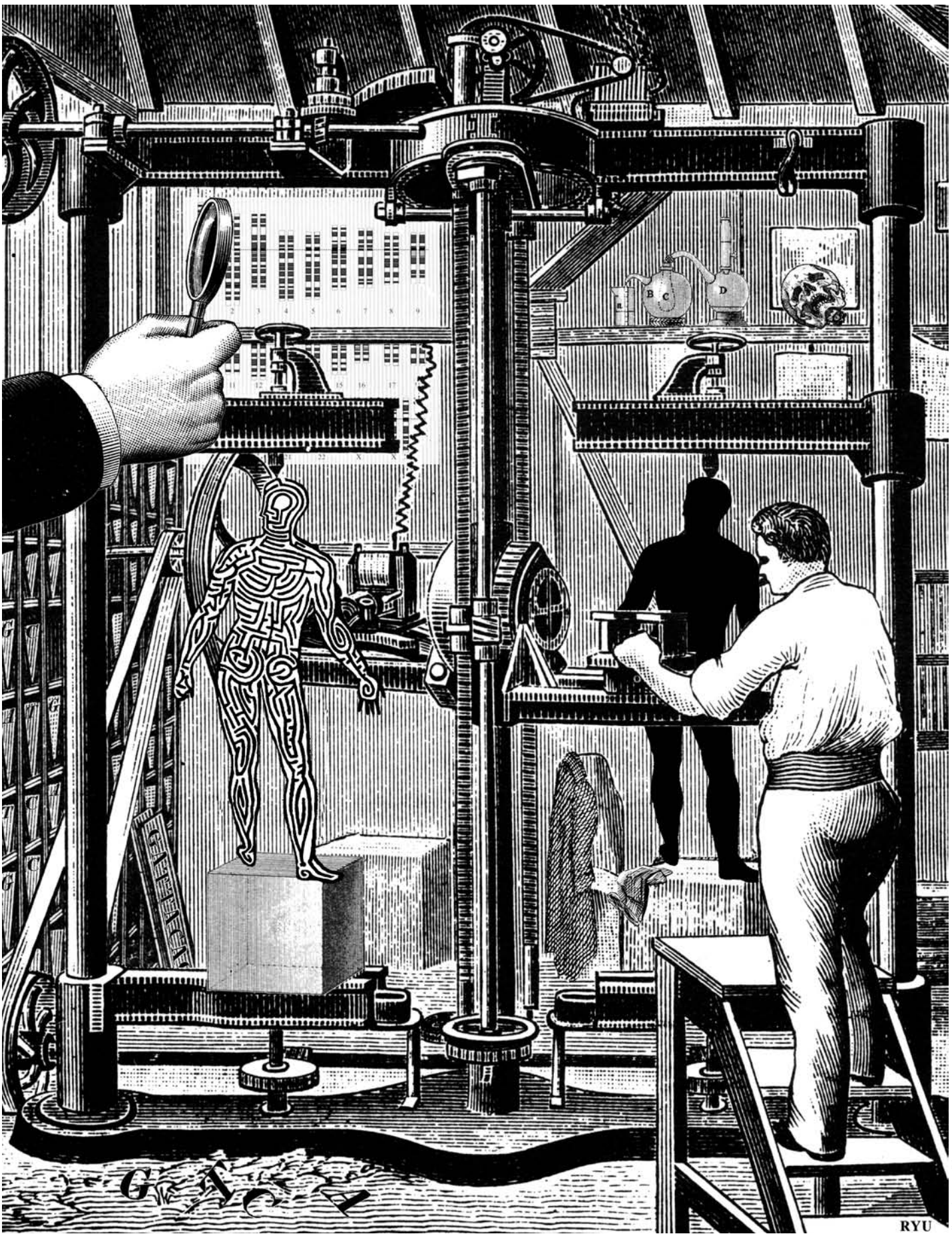
DNA-onderzoek wordt op zeer grote schaal uitgevoerd om strafzaken mee op te helderen. Dit gebeurt onder andere door het aanleggen van een databank met DNA-profielen, waarin het (forensische) DNA-profiel van verdachten, veroordeelden en sporen wordt opgeslagen. Binnenkort gaat het DNA-onderzoek een nieuwe fase in: in bepaalde gevallen wordt het mogelijk een onbekende dader op te sporen door in de databank op zoek te gaan naar een familielid van de dader, wanneer deze zelf niet daarin aanwezig blijkt te zijn. Klaas Slooten is wiskundige en werkt bij de afdeling Humane Biologische Sporen van het Nederlands Forensisch Instituut (NFI). In dit artikel gaat hij in op enkele toepassingen van kansrekening die bij DNA-onderzoek komen kijken, met name bij zoekacties in de DNA-databank.

Sinds 1994 kent Nederland specifieke DNA-wetgeving die de mogelijkheden van DNA-onderzoek in strafzaken wettelijk vastlegt. De wetgeving houdt onder meer in dat wanneer een officier van justitie bevel geeft tot DNA-afname van een persoon, deze afname verplicht is. Oorspronkelijk betrof dit alleen zeden- en geweldsmisdrijven, inmiddels alle misdrijven die worden beschreven in artikel 67 van het wetboek voor strafvordering (misdrijven waarvoor voorlopige hechtenis is toegestaan). Als gevolg hiervan beschikt de overheid over steeds meer DNA-profielen van verdachten, veroordeelden en sporen. In 1997 is de DNA-databank voor strafzaken geopend. Inmiddels (februari 2012) telt deze databank ruim 130.000 DNA-profielen van verdachten en veroordeelden, en zo'n 50.000 DNA-profielen van sporen (voor actuele cijfers zie www.forensischinstituut.nl/dna-databank).

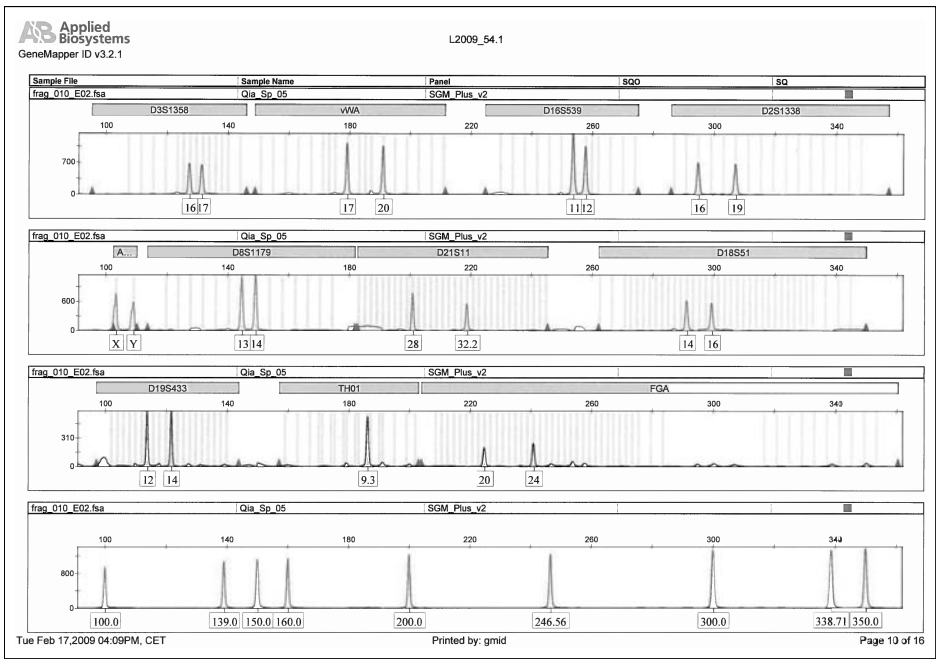
Zoals deze aantallen al aangeven is DNA-onderzoek uitgegroeid tot een op zeer grote schaal toegepast soort onderzoek: het NFI heeft in de periode 1997–2010 ruim 500.000 DNA-analyses uitgevoerd. De grote meerderheid hiervan betreft DNA-onderzoek in strafzaken, maar DNA wordt ook gebruikt voor de identificatie van slachtoffers, het verifiëren van familierelaties in immigratiezaken, et cetera.

In dit artikel zal ik ingaan op enkele wiskundige facetten van DNA-onderzoek. Hierbij zal ik vooral verslag doen van wat bekend staat als de DNA-databankcontroverse, maar ook kort op enkele andere onderwerpen ingaan.

Eerst zal ik in een notendop samenvatten wat een forensisch DNA-profiel eigenlijk is. Het DNA is, zoals welbekend, onder meer opgeslagen in de celkern in de vorm van chromosomen. Er zijn 23 chromosomen, en van elk chromosoom hebben bijna alle lichaamscellen er twee: de ene is doorgegeven door de vader, de andere door de moeder. Er zijn enkele uitzonderingen, geslachtscellen bijvoorbeeld hebben van elk chromosoom er slechts een, en rode bloedcellen hebben helemaal geen celkern, dus ook geen chromosomen. De geslachtschromosomen worden X en Y genoemd (mannen hebben XY en vrouwen XX), en alle 22 andere chromosomen heten autosomaal. Voor het bepalen van iemands forensische DNA-profiel wordt er, behalve naar het geslacht, gekeken naar een aantal lokaties op het autosomale DNA (loci genoemd) waarvan bekend is dat zich daar een stuk van variabele lengte bevindt, bestaande uit een aantal herhalingen van hetzelfde woord, bijvoorbeeld ATCTATCTATCTATCT, vier herhalingen van ATCT. Zo'n aantal herhalingen dat zich voor kan doen wordt een allel genoemd, het allel met k herhalingen noemt men simpelweg allel k, in het bijzonder zou het zojuist genoemde voorbeeld allel 4 zijn. De forensische loci hebben doorgaans tien à twintig allelen en zijn daarmee sterk onderscheidend. Omdat iedereen van elk autosomaal chromosoom er twee heeft bestaat iemands DNA-profiel op elk locus uit twee getallen, de namen van de betreffende allelen. Een forensisch DNA-profiel op k autosomale loci is dus een rijtje van $2k$ getallen. Dit maakt zo'n profiel zeer geschikt voor opname in databanken en voor onderlinge vergelijking. In het tegenwoordige standaardonderzoek is $k = 15$, en wordt zoals gezegd ook het geslacht getypeerd. Het DNA-profiel wordt bepaald door het DNA in een chemische reactie (PCR genaamd) te vermenigvuldigen om er voldoende grote hoeveelheden van te verkrijgen. Hierbij wordt een aantal cycli doorlopen (29 in



Illustratie: Ryu Tajiri



Figuur 1 Het DNA-profiel van de auteur

het standaard onderzoek) waarin in principe telkens van elk allel een getrouwe kopie gemaakt wordt. Na afloop van de PCR wordt het DNA-profiel bepaald door de allelen er in op te nemen die voldoende vaak aanwezig zijn in het resulterende PCR-produkt. Een DNA-profiel wordt door de computer gegenereerd en grafisch gerepresenteerd als een piekenpatroon, waarbij de piekhoogte een maat is voor de hoeveelheid DNA, en de positie van de piek met een bepaald allel correspondeert. In Figuur 1 staat een voorbeeld van een DNA-profiel, namelijk dat van de auteur. Merk op dat op het locus TH01 slechts één allel (namelijk 9.3) is benoemd: op dit locus zijn de twee allelen op de verschillende chromosomen gelijk aan elkaar. Overigens betekent de benaming 9.3 dat er sprake is van negen volledige herhalingen en één gedeeltelijke waar maar drie (van de vier) letters aanwezig zijn. Het getoonde DNA-profiel is verouderd in de zin dat het is vervaardigd met de SGMPlus-kit die tien autosomale loci en het geslacht typeert. Tegenwoordig worden er zoals eerder vermeld vijftien autosomale loci getypeerd, waaronder de tien van SGMPlus. Dit gebeurt met de NGM-kit. Een overzicht van de ligging op de chromosomen van de loci in deze kit en het locus SE33 (dat alleen in speciale gevallen wordt onderzocht) wordt gegeven in Figuur 2.

De autosomale loci liggen in het DNA dat niet tot expressie komt (dat wil zeggen, niet codeert voor eiwitten), zodat er geen informatie over iemands gezondheid of uiterlijk wordt verkregen. Daarnaast worden de loci zo gekozen dat ze zo statistisch onafhankelijk mogelijk zijn. Dit kan men realiseren door de loci op verschillende chromosomen te kiezen, maar omdat er voorafgaand aan de vorming van geslachtscellen recombinatie optreedt, kunnen ook loci op hetzelfde chromosoom zich onafhankelijk gedragen.

Bewijswaarde

Een forensisch laboratorium wordt gevraagd om inzicht te geven in hoeverre bewijsmateriaal wijst op de waarheid van een of andere hypothese, bijvoorbeeld de hypothese dat een kogel uit een wapen komt, een vingerafdruk van een vinger, of in het geval van DNA, een spoor (mede) van een persoon.

Voor de opkomst van DNA-onderzoek gebeurde dat door het doen van (semi-)categorische uitspraken: de forensische expert concludeert bijvoorbeeld dat een vingerafdruk afkomstig is van een bepaalde vinger, met uitsluiting van alle andere; of juist dat een vingerafdruk niet afkomstig kan zijn van een bepaalde vinger. Een light-variant is om met aan zekerheid grenzende waarschijnlijkheid te concluderen. Tegenwoordig is het inzicht ontstaan dat dit soort uitspraken, zeker als het gaat om het concluderen dat een spoor alleen afkomstig kan zijn van een bepaalde bron met uitsluiting van alle andere (bekend of niet), doorgaans wetenschappelijk niet houdbaar zijn. Hierbij heeft de opkomst van DNA-onderzoek, dat van meet af aan een solide wetenschappelijke basis had, zeker een rol gespeeld. Er wordt nu door de meeste forensische disciplines overgestapt naar wat bij DNA-onderzoek de standaardmanier van rapporteren is, soms de Bayesiaanse methode genoemd. Volgens deze werkwijze worden twee hypotheses geformuleerd, die vaak als H_p (prosecution) en H_d (defense) worden aangeduid, en berekent men de likelihood ratio (LR)

$$\frac{P(E \mid H_p, I)}{P(E \mid H_d, I)}, \qquad (1)$$

waarbij I achtergrondinformatie voorstelt. Het idee hierachter is dat de forensisch deskundige, door alleen de LR te berekenen, zich afzijdig houdt van conclusies over welke hypothese waar is, maar alleen de relatieve steun voor H_p berekent, vergeleken met die voor H_d. Het is dan aan de rechtbank om de berekening

$$\frac{P(H_p \mid E, I)}{P(H_d \mid E, I)} = \frac{P(E \mid H_p, I)}{P(E \mid H_d, I)} \frac{P(H_p \mid I)}{P(H_d \mid I)} \qquad (2)$$

uit te voeren. De rechter moet dus zelf, op grond van de achtergrondinformatie I de 'a priori' kansen $P(H_p|I)$ en $P(H_d|I)$ inschatten (dat wil zeggen, hoe kansrijk deze hypotheses zijn als het bewijs E niet bekend is), en door middel van de door de deskundige aangeleverde LR komen tot een nieuwe ('a posteriori') kansverhouding $P(H_p \mid E, I)/P(H_d \mid E, I)$ waarin het bewijs verdisconteerd is. Op die manier ontstaat er onderscheid tussen hoe sterk de 'zaak' is (gegeven door de a posteriori kansverhouding) en hoe sterk het bewijs E is (gemeten door de LR). In het bijzonder kan hetzelfde bewijs in verschillende zaken aanleiding geven tot andere overtuigingen van de rechter.

Complicaties

Wiskundig is er op het bovenstaande natuurlijk niets aan te merken, maar een natuurlijk kader voor de evaluatie van bewijs levert dit pas op als er telkens voor H_p, H_d, E en I een natuurlijke keuze is, want alleen dan is ook de bewijswaarde van E (dat wil zeggen de likelihood ratio (1), die al deze elementen bevat) ondubbelzinnig.

Dat is niet altijd zo, zoals ik dadelijk zal illustreren, en bijgevolg is er niet altijd zoiets als 'de' bewijswaarde van E. Andere misverstanden zijn dat de deskundige nooit de a priori kansen hoeft te kennen, en dat de rechter uit de a posteriori kansverhouding van H_p ten opzichte van H_d ook de a posteriori kans $P(H_p \mid E, I)$ kan berekenen. Dat laatste kan natuurlijk alleen als er geen andere hypotheses in het spel zijn.

Verder zijn er nog kwesties als hoe deze kansen eigenlijk geïnterpreteerd moeten wor-

den (frequentistisch?, als subjectieve kansen?), en is het gebruik van de benodigde kansrekening in de rechtszaal, met name bij juryrechtspraak, niet overal zonder meer mogelijk. Hier zal ik niet verder op ingaan, maar het zijn eveneens heikele punten.

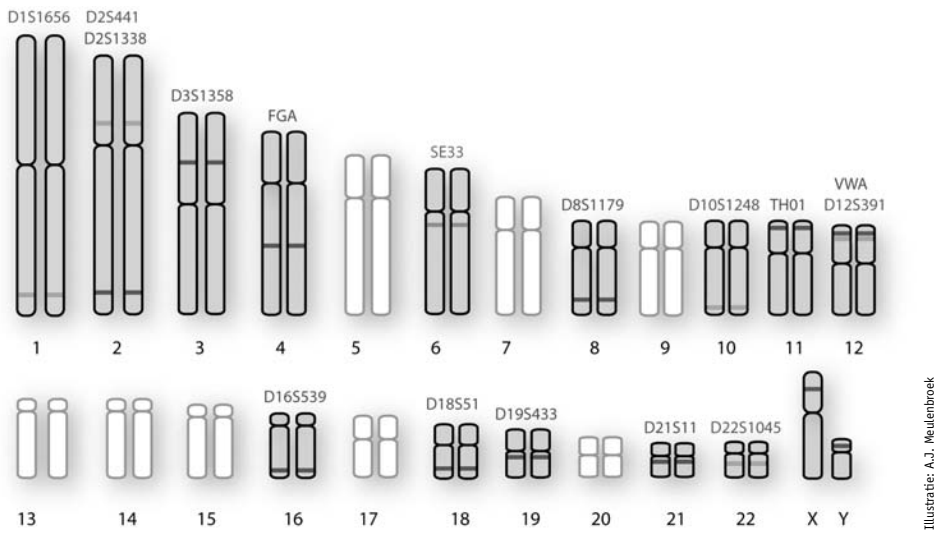
Keuze van hypotheses

Om een LR te kunnen berekenen moeten er hypotheses gekozen worden. Zowel voor H_p als H_d zijn soms meerdere zinvolle keuzes mogelijk. Het geval van een match met een DNA-profiel in de databank, dat ik zo dadelijk zal bespreken, is hier een berucht voorbeeld van, omdat deze keuzes tot zeer verschillende LR's leiden. In dit databankgeval kan je namelijk verschillende hypotheses voor H_p kiezen, waarop de kans even groot wordt wanneer er op E wordt geconditioneerd. Iets soortgelijks doet zich voor (zie bijvoorbeeld [15]) bij de keuze van H_p wanneer een verdachte matcht met één van verschillende sporen in dezelfde zaak, of met een DNA-mengprofiel.

Ook de keuze van H_d is niet altijd eenvoudig, nog afgezien van het feit dat je je kan afvragen of het formuleren hiervan niet op gespannen voet staat met het zwijgrecht: een verdachte hoeft niet uit te leggen hoe het bewijs volgens hem tot stand is gekomen. In een poging de zaak eenvoudig te houden kan men als alternatieve hypothese H_d simpelweg de ontkenning van H_p nemen. Dit vergt dan wel een model waarin $P(E \mid H_d)$ berekenbaar is, en zo'n model is meestal niet geformuleerd. Laten we weer het voorbeeld bekijken van een match tussen de DNA-profielen van een spoor en een verdachte, zeg Piet. Als Piet niet de werkelijke donor is van het spoor, hangt de kans dat hij per toeval dit DNA-profiel heeft af van zijn relatie tot die werkelijke donor: niet verwanten hebben een zeer kleine kans om per toeval hetzelfde profiel te hebben, broers hebben een veel grotere kans, en eeneïge tweelingbroers hebben dat zeker. De kans $P(E \mid H_d)$ kan dus pas berekend worden, wanneer H_d beschrijft met welke kans de werkelijke donor van het spoor welke relatie tot Piet heeft. Dit is niet realistisch en in de praktijk wordt dan ook standaard als alternatieve hypothese genomen dat een aan de verdachte niet verwante persoon de donor van het spoor is. Waakzaamheid is dus geboden: de aldus berekende LR drukt uit hoe goed Piet kan worden onderscheiden van de algemene bevolking, maar niet van zijn verwanten.

Keuze van het bewijs

Een hieraan verwant probleem is wat men tot bewijs E en wat men tot achtergrondinformatie I rekent. Ook hier volstaat de eenvoudige situatie van een DNA-match om het verschijnsel te illustreren. Stel dat verdachte S en het spoor, afkomstig van dader C, DNA-profiel g hebben, en we nemen als hypotheses $S = C$ en $S \neq C$. Wat is dan precies het bewijs? Is dat de hele collectie van feiten, dus (E_C) dat het spoor DNA-profiel g heeft, en (E_S) dat verdachte DNA-profiel g heeft? Of rekenen we E_C tot de achtergrondinformatie I en bestaat het bewijs alleen uit E_S?



Illustratie: A.J. Meulenbroek

Figuur 2 Locaties van de loci in de NGM-kit en locus SE33. De SGMPlus-loci zijn donkergrijs aangegeven.

Als de bevolking bestaat uit meerdere subpopulaties die elk hun eigen frequentie hebben van dit profiel, zullen de likelihood ratios verschillen. De eerste is gegeven door

$$\frac{P(E_C, E_S \mid S = C)}{P(E_C, E_S \mid S \neq C)} = \frac{P(E_C \mid E_S, S = C)}{P(E_C \mid E_S, S \neq C)},$$

de tweede door

$$\frac{P(E_S \mid E_C, S = C)}{P(E_S \mid E_C, S \neq C)}.$$

Ze corresponderen met de twee verschillende vragen: Hoe groot is de kans dat het spoor DNA-profiel g heeft, als de verdachte het niet heeft achtergelaten? En: Hoe groot is de kans dat de verdachte dit DNA-profiel heeft, als hij het spoor niet heeft achtergelaten? Voor de laatste vraag hoef je (verwantschap negerend) alleen de frequentie van het DNA-profiel in de subpopulatie van de verdachte te weten, maar voor de eerste moet je de kans π_i kennen dat het spoor is achtergelaten door iemand uit de i-de subpopulatie. In dit geval hangt de likelihood ratio van die a priori kansen π_i af!

Ook hier geldt dat de a posteriori kansverhouding niet van deze keuzes afhangt, die alleen tot gevolg hebben dat er een factor uitgewisseld wordt tussen de likelihood ratio en de prior odds. In [19] hebben Ronald Meester en ik een uitgebreide studie gemaakt van dit verschijnsel, waarbij we ook onderzoeken wat de invloed is van onzekerheid over de bevolkingsfrequenties van de DNA-profielen, en over tot welke subpopulatie iemand behoort.

Databanken

Directe vergelijkingen

Eerst gaan we terug naar de discussie die in de jaren negentig werd gevoerd over de bewijswaarde van een match in de databank. De vraag rees hoe die zou verschillen van de bewijswaarde van een enkelvoudige spoor-persoonvergelijking, als gevolg van het grote aantallen vergelijkingen dat was uitgevoerd. Laten we S de persoon noemen die als enige matcht in de databank $\mathcal{D}$, C de werkelijke donor van het spoor, en N de grootte van $\mathcal{D}$. Als we werken in een homogene populatie waarin p de frequentie van het betreffende profiel is, en we negeren verwantschap, dan doen de hierboven genoemde complicaties zich niet voor en is het eenvoudig in te zien dat de likelihood ratio voor de hypotheses $S = C$ versus $S \neq C$ geassocieerd met enkel de match tussen de twee profielen van S en C, gelijk is aan $1/p$. De vraag was nu hoe te verrekenen dat de match afkomstig was uit $\mathcal{D}$, dat wil zeggen dat er ook $N - 1$ vergelijkingen waren uitgevoerd waarin geen match was gevonden. Maakt dit de match nu sterker of zwakker? Aan de ene kant geldt natuurlijk nog steeds dat als S niet gelijk is aan C, hij een kans p heeft om met het spoor te matchen. Aan de andere kant werd er geredeneerd dat, omdat *elke* match tot een verdenking zou hebben geleid,

LR	$S = C$	$C \in \mathcal{D}$
$S \neq C$	$\frac{1}{p}\frac{1-\pi_1}{1-\pi_\mathcal{D}}$	$\frac{1}{p}\frac{\pi_1(1-\pi_1)}{\pi_\mathcal{D}(1-\pi_\mathcal{D})}$
$C \notin \mathcal{D}$	$\frac{1}{p}$	$\frac{1}{p}\frac{\pi_1}{\pi_\mathcal{D}}$

Tabel 1 LR's voor een databankmatch

de relevante kans niet langer de kans is dat S bij toeval matcht, maar dat er iemand in de hele databank per toeval matcht. Die kans is $1-(1-p)^N$, hetgeen meestal zeer goed wordt benaderd door Np.

De Amerikaanse National Research Council gaf in haar rapport van 1992 (zie [6]) de aanbeveling om de geconstateerde match in de databank geheel als bewijs achterwege te laten, en in plaats daarvan voor zowel spoor als verdachte aanvullende DNA-profielen (op andere loci) te genereren. Als de match dan stand zou houden, zou de bewijswaarde alleen betrekking moeten hebben op deze loci. Nog afgezien van de theoretische bezwaren tegen dit wel heel conservatieve advies is er de praktische complicatie dat additionele loci typeren niet altijd mogelijk is, wat dan zou betekenen dat een match met zo'n spoor zonder gevolgen zou blijven.

De redenering die leidt tot het berekenen van de kans dat er iemand in de databank ten onrechte matcht, is gestoeld op principes van de frequentistische statistiek. Intussen ontstond zoals gezegd echter consensus dat de bewijswaarde gerepresenteerd zou moeten worden door een likelihood ratio. Daarmee was de discussie zeker niet beslecht. Hiervoor zijn immers meerdere mogelijkheden (zie hierboven): we kunnen als eerste hypothese $S = C$ of $C \in \mathcal{D}$ nemen, en als tweede kunnen we $S \neq C$ en $C \notin \mathcal{D}$ nemen. Dit levert vier verschillende likelihood ratios. De twee die aanleiding gaven tot de discussie waren de paren $S = C$ versus $S \neq C$ en $C \in \mathcal{D}$ versus $C \notin \mathcal{D}$.

Wie kiest voor $S = C$ versus $S \neq C$, concludeert [1–2, 8] dat de bewijswaarde van een DNA-databankmatch groter is dan $1/p$, omdat de $N-1$ uitsluitingen de kans op $S = C$ alleen maar vergroten. Omdat $S = C$ ook is wat een rechtbank wil onderzoeken, werd dit door velen als een juiste weergave van de bewijswaarde gezien. Zoals Balding en Donnelly aanvoeren in [2] als bezwaar tegen de hypotheses $C \in \mathcal{D}$ versus $C \notin \mathcal{D}$: "A court is not concerned with the collective guilt or innocence of the database."

Aan de andere kant maakten sommigen bezwaar (zie [20]) tegen het nemen van $S = C$ als hypothese omdat deze wordt geformuleerd nadat de match is gevonden, en dus een data-afhankelijke hypothese zou zijn waardoor de bewijswaarde geflatteerd wordt. Immers, zo betogen zij, als $p = 1/N$, dan is de kans op een unieke match even groot wanneer $C \in \mathcal{D}$ als wanneer $C \notin \mathcal{D}$, en is dat te verenigen met eraan toekennen van een bewijswaarde van $1/p$ of meer? In plaats daarvoor wordt voorgesteld om de hypotheses $C \in \mathcal{D}$ en $C \notin \mathcal{D}$ te nemen, hetgeen een likelihood ratio van $1/(Np)$ oplevert, wanneer de a priori kansen $P(C = d_i)$ gelijk zijn. Dit sluit aan bij het in 1996 door de Amerikaanse National Research Council (NRC) gepubliceerde rapport waarin werd aanbevolen om $1/(Np)$ te nemen "if one wishes to describe the impact of the DNA evidence under the hypothesis that the source of the evidence sample is someone in the database" [17].

Het probleem is wederom (zoals ook geconcludeerd in [15]) dat het concept bewijswaarde tot op zekere hoogte arbitrair is, en alleen de a posteriori kansverhouding goed gedefinieerd is. Tegenover het feit dat $S = C$ een data-afhankelijke hypothese is, staat dat *alle* hypotheses $C = d_i$ (voor $d_i \in \mathcal{D}$) voorafgaand aan de zoekactie gedefinieerd zijn, anders zouden we deze niet uitvoeren. Het is alleen zo dat na afloop de hypothese $S = C$ als enige nog mogelijk is.

Aan de andere kant is er tegen Balding en Donnelly in te brengen dat de rechtbank in het geval van een unieke match wel degelijk geïnteresseerd is in de collectieve kans op schuld of onschuld van de databank. Sterker nog, dat is precies de kans die de rechtbank wil berekenen: immers $P(C \in \mathcal{D} \mid E) = P(S = C \mid E)$.

Een overzicht van de likelihood ratios die in dit verband verkregen worden, is te zien in Tabel 1, met $\pi_1 = P(S = C)$ en $\pi_\mathcal{D} = P(C \in \mathcal{D})$. Merk op dat al deze likelihood ratios, met uitzondering van die voor $S = C$ versus $C \notin \mathcal{D}$, van de a priori kansen afhangen. De a posteriori odds zijn steeds hetzelfde en gelijk aan $P(S = C \mid E)/P(S \neq C \mid E) = \pi_1/(p(1-\pi_\mathcal{D}))$. Het is dus simpelweg een kwestie van langs welke weg men de kans op schuld wil berekenen: een andere keuze leidt er alleen toe dat de likelihood ratio en de a priori odds een factor uitwisselen. Wie $S = C$ als eerste hypothese neemt, komt op een hogere bewijswaarde ($1/p$ of meer), en het risico is dat uit het oog wordt verloren hoe extreem klein de a priori kansen kunnen zijn. Immers, in een grote databank kunnen niet veel mensen een hoge a priori kans hebben om C te zijn, gemiddeld komt die hoogstens op $1/N$ uit.

Wie daarentegen $C \in \mathcal{D}$ neemt, komt op een kleinere bewijswaarde, die zelfs in het nadeel van $C \in \mathcal{D}$ kan uitvallen. Bij uniforme prior en alternatief $C \notin \mathcal{D}$ gebeurt dit als $Np > 1$. Het risico is dat nu uit het oog wordt verloren dat, alhoewel de databank minder matches opleverde dan je zou verwachten als de dader erin zat, de kans op $S = C$ toch wel degelijk behoorlijk is toegenomen. Daarnaast moet voor deze aanpak $P(C \in \mathcal{D})$ worden berekend, hetgeen ook niet zonder complicaties is.

De discussie laait af en toe nog steeds op. Zo publiceerde de German Stain Commission (GSC) onlangs nog aanbevelingen [10] waarin een voorkeur voor $1/(Np)$ wordt uitgesproken als meest relevante statistiek. Dit standpunt werd onder vuur genomen in een reactie van Biedermann en Taroni [5, 21], maar de GSC bleef bij haar standpunt [13]. In de discussie kwamen de gebruikelijke, hierboven besproken, argumenten weer aan bod.

Er is, mijns inziens, echter nog een andere kans die interessant is om te bepalen: de kans dat, gegeven dat de databank een unieke match oplevert (maar zonder te weten met wie), dat die match met de juiste persoon is. We kunnen dit de effectiviteit van de databank noemen, want het beschrijft hoe 'betrouwbaar' de matches erin zijn. Laten we E_1 schrijven voor de gebeurtenis dat er een unieke match is, S voor de persoon waar dat mee is en p de populatiefrequentie van het profiel in kwestie. Dan geldt (zie [19]), ongeacht de a priori kansen voor de verdeling van C op $\mathcal{D}$, dat

$$\frac{P(S = C \mid E_1)}{P(S \neq C \mid E_1)} = \frac{1}{Np}\frac{P(C \in \mathcal{D})}{P(C \notin \mathcal{D})}, \qquad (3)$$

Nu moeten we dus $P(C \in \mathcal{D})$ kennen. In de literatuur neemt men hiervoor doorgaans de fractie van de bevolking die in de databank zit. In dat geval is, als de bevolking grootte n heeft, (3) gelijk aan $1/(p(n-N))$ hetgeen toeneemt met toenemende N. Daaruit volgt dus dat hoe groter de databank is, hoe groter diens effectiviteit is. Echter, de aanname dat $P(C \in \mathcal{D})$ gelijk is aan de bevolkingsfractie is natuurlijk buitengewoon onrealistisch. Immers, in de Nederlandse databank heeft tot nu toe (zie [4]) zo'n 44% van de ooit opgenomen spoorprofielen een match met een persoon opgeleverd!

Stel dus dat we nu een andere uitdrukking voor $P(C \in \mathcal{D})$ nemen, bijvoorbeeld $P(C \in \mathcal{D}) = \sqrt{N/n}$; dan is dus bijvoorbeeld in een databank die 1% van de bevolking bevat, de kans op $C \in \mathcal{D}$ gelijk aan 10%. In dat geval kan je laten zien dat de kansverhouding (3) *daalt* wanneer N toeneemt van 1 naar

$n/4$, daar een minimum heeft en dan weer monotoon stijgt (naar oneindig als $N = n$). Het is dan dus niet zo dat een grotere databank effectiever is. Als de databank groeit, groeit ook de kans dat de dader erin wordt opgenomen, maar eveneens groeit de kans dat er per toeval iemand met hetzelfde profiel in terechtkomt. Welk effect de overhand heeft, hangt van alle factoren in (3) af. Het model $P(C \in \mathcal{D}) = \sqrt{N/n}$ is natuurlijk ook niet realistisch, maar lijkt wel realistischer dan $P(C \in \mathcal{D}) = N/n$.

Tot slot van deze discussie nog even de vraag: valt deze exercitie onder de recreatieve kansrekening, of maakt het in de praktijk iets uit welke likelihood ratio men als uitgangspunt neemt? In principe natuurlijk niet, omdat een rechtbank over een schuldvraag zou moeten beslissen op basis van hoe sterk de zaak als geheel is, in plaats van alleen op basis van hoe sterk het bewijs is. In de praktijk kan de verleiding natuurlijk groot zijn om de overtuiging over hoe sterk de zaak is teveel, of zelfs geheel, te baseren op hoe sterk het bewijs is. Voor volledige DNA-profielen op 15 loci doet het er weinig toe: dan is de populatiefrequentie p in de orde van 10^{-20}, en zelfs met een databankgrootte N van enkele miljoenen, is Np nog extreem klein. Er is dus hoe dan ook sprake van zeer sterk bewijs.

Maar niet alle profielen zijn volledig, en niet alle landen typeren dezelfde loci. Tussen de Nederlandse en verschillende buitenlandse DNA-databanken worden sinds 2008 geautomatiseerd profielen uitgewisseld en vergeleken (in het kader van het zogenaamde verdrag van Prüm). De vergelijking met de Duitse DNA-databanken is bijzonder omdat dan doorgaans slechts zes of zeven loci kunnen worden vergeleken, aangezien beide landen andere verzamelingen loci in het standaard DNA-onderzoek toepassen. Voor een profiel van zes loci is p in de orde van 1 op enkele tientallen miljoenen, toch blijkt (zie [3]) uit nader onderzoek dat twee van de drie matches van zulke profielen per toeval zijn. Hier hebben we een voorbeeld van zeer kleine a priori kansen, die ondanks een forse bewijswaarde niet tot een grote a posteriori kans leiden. Daarom wordt er in geval van zo'n match standaard eerst een profiel uitgebreid zodat meer loci vergeleken kunnen worden.

Ook worden er DNA-mengprofielen opgenomen in de databank, die afkomstig zijn van meerdere personen. Deze kunnen zijn opgeslagen door alle getypeerde allelen op te nemen, zonder indicatie welke allelen van dezelfde donor zijn. In dat geval komt iemand in aanmerking als donor als die uitsluitend allelen heeft die in het mengsel voorkomen, en de kans p daarop is dusdanig dat $Np > 1$ zeker mogelijk is. In dat geval is een goed begrip van de discussie hierboven dus zeker belangrijk.

Foto: Beeldbank NFI

Figuur 3 DNA-onderzoek bij het NFI

Het NFI neemt in het geval van een match die uit de databank wordt verkregen, in het rapport een kader op met als titel 'Aandachtspunt bij een DNA-databankmatch'. Hierin wordt de lezer van het rapport erop gewezen dat: "Naarmate het aantal DNA-profielen in de DNA-databank toeneemt, neemt ook de kans toe dat bij een zoekactie in de DNA-databank een match wordt verkregen met een persoon van wie het onderzochte sporenmateriaal niet afkomstig is. [...] Met deze mogelijkheid moet met name rekening worden gehouden wanneer het een DNA-databankmatch betreft met een spoor waarvan een onvolledig DNA-profiel of een DNA-mengprofiel is verkregen. [...]"

Familial searching

Natuurlijk levert de databank niet altijd een match op. Dat betekent vooralsnog dat de identiteit van de dader onbekend blijft na de zoekactie (er wordt natuurlijk nog wel gezocht tegen de profielen die later worden opgenomen), maar dat gaat zeer waarschijnlijk binnenkort veranderen. Immers, DNA erft over van ouder naar kind, en daarom lijken de DNA-profielen van naaste verwanten doorgaans meer op elkaar dan die van niet (of nauwelijks) verwante personen. Op basis hiervan is er momenteel een wetswijziging in voorbereiding (en reeds door Tweede en Eerste Kamer aangenomen) die het gebruik van de databank drastisch zal verruimen: als de (vermeende) dader niet in de databank blijkt te zijn opgenomen, mag er in bepaalde gevallen ook worden gezocht naar mogelijke verwanten van hem (of haar) in de databank. Op die manier kan de identiteit van de dader wellicht alsnog worden achterhaald. Dit proces wordt *familial searching* genoemd. In het Verenigd Koninkrijk zijn met deze techniek sinds de invoering ervan in 2004 enkele tientallen zaken opgelost. Verder zijn er wereldwijd nog slechts enkele andere landen of staten die dit soort onderzoek uitvoeren.

Hoe gaat zo'n zoekactie in zijn werk? Stel dat de databank is $\mathcal{D} = (d_1, \ldots, d_N)$, en het doelprofiel is C, en we kiezen een vorm van verwantschap (in de praktijk is alleen ouder/kind of broer/zus werkbaar). Het doelprofiel C wordt dan vergeleken met elk van de databankprofielen d_i. De genetische aanwijzing voor verwantschap van de gekozen vorm tussen C en d_i wordt gegeven door (alweer) een likelihood ratio die uitdrukt hoeveel waarschijnlijker het is deze twee profielen aan te treffen onder de hypothese van deze vorm van verwantschap dan onder de hypothese van niet-verwantschap. We verkrijgen dus de vector $\mathbf{r} = (r_1, \ldots, r_N)$, met r_i de likelihood ratios tussen C en d_i die deze verwantschap afzetten tegen niet-verwantschap. De meeste r_i zijn klein omdat de meeste C en d_i heel weinig op verwanten van elkaar lijken, maar het is eveneens mogelijk dat een aantal r_i groot zijn omdat C en d_i per toeval veel op verwanten lijken.

Enkele elementen van de zojuist beschreven databankcontroverse zijn dus weer aan-

wezig: een groeiende databank bevat met steeds grotere kans een verwant van C, maar bevat ook met steeds grotere kans profielen die geheel toevallig op verwanten van C lijken. Daar staat tegenover dat er nu geen sprake is van geheel overeenkomende profielen maar van een gedeeltelijke overeenkomst, met een corresponderende likelihood ratio die zeer veel waarden aan kan nemen.

Wellicht omdat familial searching nog maar vrij kort wordt toegepast en het aantal landen dat het uitvoert beperkt is, is de literatuur erover niet heel uitgebreid. Het meeste gepubliceerde onderzoek is (al dan niet gesimuleerd) empirisch, waarbij wordt beschreven hoe hoog een verwant in een databank scoort indien die op likelihood ratio (of aantal gemeenschappelijke allelen) geordend wordt (zie [7, 9, 14, 16]. Daarnaast ontstaat weer, omdat het risico op toevalsmatches reëel is, discussie over de interpretatie van de zoekresultaten. In [18] wordt geadviseerd om de likelihood ratio r_i te delen door het aantal leden van de databank en pas verder onderzoek in te stellen als dit quotient r_i/N voldoende groot is, hetgeen sterk doet denken aan de reeds besproken discussie over de bewijswaarde van directe matches.

Het model van hierboven waarin de klassieke databank-controverse kan worden begrepen, hebben Ronald Meester en ik [11–12] onlangs uitgebreid naar deze meer algemene situatie. In ons model gaan we er van uit dat de databank ofwel geen, ofwel één verwant bevat van het gegeven type, die noemen we R. Dan kan je laten zien dat, indien $\pi_i = P(R = i)$ de a priori kans voorstelt dat de verwant databanklid i is,

$$P(R = i \mid \mathbf{r}) = \frac{r_i \pi_i}{\sum_{k=1}^{N} r_k \pi_k + P(R \notin \mathcal{D})}.$$

De corresponderende likelihood ratio, die ik in het vervolg databank-likelihood ratio zal noemen, is

$$\frac{P(\mathbf{r} \mid R = i)}{P(\mathbf{r} \mid R \neq i)} = \frac{r_i}{\sum_{k=1, k\neq i}^{N} r_k \frac{\pi_k}{1-\pi_i} + \frac{P(R \notin \mathcal{D})}{1-\pi_i}}. \quad (4)$$

Merk op dat deze een functie is van de a priori kansen π_k, en de individuele likelihood ratios r_k.

De databank likelihood ratio in het voordeel van $R \in \mathcal{D}$ wordt gegeven door

$$\frac{P(\mathbf{r} \mid R \in \mathcal{D})}{P(\mathbf{r} \mid R \notin \mathcal{D})} = \frac{\sum_{k=1}^{N} r_k \pi_k}{P(R \in \mathcal{D})}. \quad (5)$$

Als de verdeling van R op $\mathcal{D}$ uniform is, dat wil zeggen $P(R = i) = P(R \in \mathcal{D})/N$, dan reduceert (5) tot

$$\frac{\sum_{k=1}^{N} r_k}{N},$$

hetgeen betekent dat de bewijswaarde van de gehele vector **r** van zoekresultaten in het voordeel van $R \in \mathcal{D}$, gelijk is aan de gemiddelde bewijswaarde voor $R = d_i$. Ook hier kan het weer gebeuren dat **r** als geheel de kans op $R \in \mathcal{D}$ verkleint, terwijl er toch voor sommige d_i zeer sterke aanwijzingen zijn dat $R = i$.

Het is niet moeilijk in te zien dat de verwachtingswaarde van r_i, als in werkelijkheid $R \neq i$, gelijk is aan 1. Dit betekent dat als $R \notin \mathcal{D}$, we verwachten dat $r_1 + \cdots + r_N = N$. Met andere woorden, gegeven welke persoon dan ook is het te verwachten dat de databank een aantal leden telt die genetisch erg op verwanten van die persoon lijken. Maar pas als de totale likelihood ratio groter is dan het aantal leden van de databank is er sprake van een aanwijzing dat de databank daadwerkelijk een verwant bevat (voor uniforme prior).

Hieruit wordt ook duidelijk hoe je een zoekstrategie kan definiëren die de gezochte verwant (als die inderdaad in de databank zit) met een berekenbare kans oplevert: als $\mathcal{D}(k)$ correspondeert met de k grootste producten $r_i\pi_i$, en $\mathcal{D}^\alpha$ is de kleinste $\mathcal{D}(k)$ zodat $\sum_{j\in\mathcal{D}(k)} r_j\pi_j \geq \alpha \sum_{j=1}^{N} r_j\pi_j$, dan is $P(R \in \mathcal{D}^\alpha \mid R \in \mathcal{D}) \geq \alpha$.

Op die manier kan je bijvoorbeeld besluiten wanneer je met de zoekactie naar een verwant wilt stoppen als die tot dusverre niets heeft opgeleverd, of welke personen je moet onderzoeken om een vooraf bepaalde kans te hebben om de verwant te vinden indien die in de databank aanwezig is.

Ook andere zoekstrategieën zijn mogelijk, en niet alle vragen over familial searching zijn hiermee beantwoord. Recent hebben Ronald Meester en ik een NWO-subsidie toegekend gekregen voor verder onderzoek. Hierin gaan we onder meer kijken naar zoekstrategieën die van het daderprofiel afhangen, en het voordeel hebben dat er geen a priori kansen nodig zijn. Verder gaan we onder meer DNA-verwantschapsonderzoek in een grootschalig bevolkingsonderzoek modelleren. Zulke onderzoeken (met deelname op vrijwillige basis) worden door de nieuwe wetgeving ook mogelijk gemaakt: als familial searching in de databank geen resultaat heeft, kan zo'n onderzoek wellicht alsnog leiden naar de dader.

Tot slot

De databankdiscussie is feitelijk wel uitgekristalliseerd. Er zijn echter nog tal van andere gebieden in het DNA-onderzoek waar wiskundig onderzoek nodig is. Op enkele daarvan zal ik tot besluit van dit artikel nog kort ingaan.

Onzekere DNA-profielen

Tot nu toe ben ik ervan uitgegaan dat er over de DNA-profielen zelf geen twijfel mogelijk is. Wat databanken betreft is dat niet zo onrealistisch omdat een profiel er pas in wordt opgenomen als de DNA-deskundige niet twijfelt aan de juistheid en volledigheid ervan. Het komt echter regelmatig voor, met name bij DNA-profielen van sporen die weinig DNA bevatten, dat het onzeker is of alle allelen aanwezig in de bemonstering gedetecteerd zijn en of alle gedetecteerde allelen ook echt in het DNA in de bemonstering zaten.

Dit komt omdat in werkelijkheid niet in elke cyclus van de PCR van elk allel een kopie wordt gemaakt: elk allel heeft een bepaalde kans (niet constant over alle cycli) om gekopieerd te worden. Daarnaast is de gemaakte kopie met een kleine kans niet gelijk aan het origineel. Een kopieerfout bestaat er meestal uit dat de kopie een herhaling minder bevat dan het origineel: de kopie van allel a is dan allel $a-1$ (dat wil zeggen een herhaling minder van het woordje op dat locus). Bij voldoende uitgangsmateriaal levert dit in de praktijk geen problemen op, maar wanneer het uitgangsmateriaal zeer weinig DNA bevat, en er tijdens de PCR van een bepaald allel te weinig worden gekopieerd, kunnen er allelen die wel in de bemonstering zitten niet benoemd worden in het DNA-profiel. Dit wordt (allelische) drop-out genoemd. Ook kunnen kopieerfouten zodanig vaak voorkomen dat een allel ten onrechte in het DNA-profiel wordt opgenomen. Vrijwel altijd wordt dan allel $a-1$ extra benoemd terwijl de bemonstering allel a bevat (dat ook is benoemd). Het lastige is natuurlijk dat aan het DNA-profiel niet altijd te zien is of alle erin benoemde allelen echt zijn, en of alle in de bemonstering aanwezige allelen benoemd zijn. In het bijzonder is het soms niet uit te sluiten dat een persoon met allelen (a, b) de donor is van een spoor waarin alleen allel a is aangetroffen. Ook het meermaals uitvoeren van een PCR op dezelfde bemonstering, het opvoeren van het aantal cycli of het inzetten van gevoeligere DNA-technologie dan het standaardonderzoek, leiden er niet altijd toe dat er een betrouwbaar DNA-profiel ontstaat. Momenteel wordt aan mogelijke matches met dit soort DNA-profielen dan ook geen getalsmatige bewijswaarde gekoppeld. Aan rekenmodellen om dit wel te kunnen doen wordt gewerkt, maar een allesomvattend model lijkt nog ver weg.

Mengsels

Wanneer het DNA van meerdere personen in een bemonstering aanwezig is, wordt het DNA-profiel een piekenpatroon met bij elk locus een wisselend aantal pieken met verschillende hoogtes. Vaak is niet te zien wat de genotypes van de verschillende donoren zijn, of zelfs maar hoeveel donoren er zijn. Dat komt omdat n personen op elk locus samen $2n$ allelen hebben, maar niet per se $2n$ verschillende allelen. Ook kan er weer dropout voorkomen, als er van sommige van de donoren zeer weinig DNA in de bemonstering zit. Wat kunnen we nu zeggen over een match tussen een persoon S en een mengsel M, dat wil zeggen de constatering dat S uitsluitend allelen heeft die ook in M voorkomen? Voor zover er al wordt gerekend aan dergelijke mengprofielen, bestaat de berekening momenteel meestal uit het uitrekenen van deze inclusiekans: de kans dat een willekeurige persoon uitsluitend allelen heeft die ook in het mengsel voorkomen. Dit is echter een statistiek van het mengsel zelf, die niet direct de bewijswaarde geeft. Inderdaad is het zo dat van iedereen die niet is uitgesloten, er mensen zijn die veel meer als donor in aanmerking komen op grond van hun genotype dan anderen, omdat de piekhoogtes in het DNA-profiel beter aansluiten bij donorschap of omdat ze de zeldzamere allelen in het mengsel verklaren. Het berekenen van een likelihood ratio zou voor veel forensische laboratoria de voorkeur hebben, maar er is geen consensus over hoe die eruit zou moeten zien. Ook hier zijn weer modellen nodig die piekhoogtes kunnen meenemen in de berekening. Daarnaast zal het resultaat ook afhangen van het aantal donoren dat elke hypothese veronderstelt: als een mengsel op een bepaald locus vier verschillende allelen $a_1, \ldots, a_4$ heeft, en S heeft op dat locus genotype (a_1, a_1), dan is S uitgesloten als donor als dit mengsel precies twee donoren heeft, maar niet als het er ook meer kunnen zijn.

Verder onderzoek

Het kunnen rekenen aan de onvolledige en mengprofielen is onder meer zo belangrijk omdat de sporen waarvan het technisch mogelijk is een DNA-profiel te vervaardigen steeds kleiner worden, tot aan enkele cellen aan toe. Dit soort sporen geeft echter veel vaker aanleiding tot de hierboven beschreven problemen dan de 'traditionele' sporen waarin veel DNA aanwezig is. Daarnaast is het van deze sporen veel minder duidelijk wat hun relatie is tot het delict. Ook dit beïnvloedt natuurlijk hun bewijswaarde.

Ik heb het in dit artikel alleen gehad over het autosomale DNA, maar ook van het Y-chromosoom en van het mitochondriale DNA kunnen DNA-profielen worden verkregen. Van deze profielen zijn de frequenties in de bevolking echter veel lastiger te bepalen, omdat ze zich als een soort genetische achternaam gedragen en veel geografische verschillen in verspreiding vertonen. Dit maakt ze meer geschikt voor een exclusie dan als kwantificeerbaar bewijsmateriaal. Daarnaast is het typeren van RNA in opkomst, omdat dit gebruikt kan worden om het type cel dat in de bemonstering is opgenomen te bepalen (is een minimaal spoor een bloedspoor?). Ook zijn er tests ontwikkeld om uiterlijk waarneembare kenmerken zoals oog- en haarkleur, leeftijd en lengte te voorspellen aan de hand van het genetische materiaal. Er zijn, kortom, in het forensische DNA-onderzoek nog volop terreinen waar wiskundigen een bijdrage kunnen leveren.

Referenties

1 D.J Balding and P. Donnelly, Inference in Forensic Identification, *Journal of the Royal Statistical Society*, Series A **158** (1995), no. 1, 21–53.

2 D.J. Balding and P. Donnelly, Evaluating DNA profile evidence when the suspect is found through a database search, *J. For. Sc.* **41** (1996), 603–607.

3 C.P. van der Beek, Forensic DNA Profiles Crossing Borders in Europe (Implementation of the Treaty of Prüm), Profiles in DNA 2011, www.promega.com/resources/articles/profiles-in-dna/2011/forensic-dna-profiles-crossing-borders-in-europe

4 C.P. van der Beek, Jaarverslag 2010 Nederlandse DNA-databank voor Strafzaken, verkrijgbaar via www.forensischinstituut.nl/dna-databank

5 A. Biedermann, S. Gittelson, and F. Taroni, Recent misconceptions about the 'database search problem': a probabilistic analysis using Bayesian networks, *Forensic Science International* **212** (2011), 51–60.

6 National Research Council, DNA technology in forensic science, *Nat. Acad. Press*, Washington, 1992.

7 J.M. Curran and J.S. Buckleton, Effectiveness of familial searches, *Science and Justice* **84** (2008), 164–7.

8 A.P. Dawid and J. Mortera, Coherent Analysis of Forensic Identification Evidence, *Journal of the Royal Statistical Society*. Series B (Methodological) **58** (1996), no. 2, 425–443.

9 J. Ge, R. Chakraborty, A. Eisenberg, et al., Comparisons of Familial DNA Database Searching Strategies, *J. For. Sciences* **56** no. 6 (2011), 1448–1456.

10 P.M. Schneider et al., Allgemeine Empfehlungen der Spurenkommission zur statistischen Bewertung von DNA-Datenbank Treffern, *Rechtsmedizin* **20** (2010), 111–115.

11 K. Slooten and R. Meester, Statistical Aspects of Familial Searching, *Forensic Science International: Genetics Supplement Series* **3**(2011), e617-e619.

12 K. Slooten and R. Meester, Database likelihood ratios and familial DNA searching, submitted.

13 R. Fimmers et al., Erwiderung zum Brief von Taroni et al. zum Beitrag von Schneider et al. "allgemeine Empfehlungen der Spurenkommission zur statistischen Bewertung von DNA-Datenbank Treffern", *Rechtsmedizin* **21** (2011), 57–60.

14 T. Hicks, F. Taroni, J. Curran, et al., Use of DNA profiles for investigation using a simulated national DNA database: Part II. Statistical and ethical considerations on familial searching, *For. Sc. Int.:Genetics* **4** (2010), no. 5, 232–8.

15 R. Meester and M. Sjerps, The Evidential Value in the DNA Database Search Controversy and the Two-Stain Problem, *Biometrics* **59** (2003), 727–732.

16 S. Myers et al., Searching for first-degree familial relationships in California's offender DNA database: Validation of a likelihood ratio-based approach, *For. Sc. Int.: Gen.* **5** (2011), no. 5, 493–500.

17 National Research Council, The evaluation of forensic DNA evidence, *Nat. Acad. Press*, Washington, 1996.

18 Scientific Working Group on DNA Analysis Methods Ad Hoc Committee on Partial Matches, SWGDAM Recommendations to the FBI Director on the "Interim Plan for the Release of Information in the Event of a 'Partial Match' at NDIS", *For. Sc. Comm.* **11** (2009), no. 4.

19 K. Slooten and R. Meester, Forensic Identification: the Island Problem and its generalizations, *Stat. Neerl.* **65** (2011), 202–237.

20 A. Stockmarr, Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search, *Biometrics* (1999), 671–677.

21 F. Taroni, A. Biedermann, R. Coquoz, T. Hicks, and C. Champod, Statistische Bewertung von DNA-Databank Treffern, *Rechtsmedizin* **21** (2011), 55–60.