

Mathisca de Gunst

Afdeling Wiskunde, Faculteit Exacte Wetenschappen

Vrije Universiteit, De Boelelaan 1081a

1081 HV Amsterdam

degunst@cs.vu.nl

Overzichtartikel

Wiskundige modellering van

Wiskunde kan een nuttige rol spelen bij het oplossen van biologische problemen. De hierbij benodigde wiskundige technieken zullen van probleem tot probleem zeer verschillen. In sommige gevallen dient zelfs nieuwe wiskunde te worden ontwikkeld. Aan de hand van twee voorbeelden wordt geïllustreerd hoe biologie op haar beurt een inspiratiebron voor wiskundigen kan zijn. Het eerste voorbeeld betreft chemische carcinogenese; het tweede de kinetiek van ionkanalen.

Dit artikel is eerder verschenen in het *Liber Amicorum* ter gelegenheid van het afscheid van K.R. Libbenga als hoogleraar in de Algemene Plantkunde aan de Universiteit Leiden op 19 november 2002. De auteur is in 1988 bij K.R. Libbenga en W.R. van Zwet gepromoveerd op onderzoek op het gebied van stochastische modellering van de groei van plantencelpopulaties.

Met de razendsnelle ontwikkelingen in de moleculaire biologie van dit moment wordt de wetenschappelijke wereld er langzamerhand van doordrongen dat de biologie niet meer zonder wiskunde en informatica kan. De roep om multidisciplinaire onderzoeksverbanden, met name tussen moleculair biologen enerzijds en wiskundigen en informatici anderzijds, wordt steeds luider. Immers, de biologische systemen die onderzocht dienen te worden, zijn van een steeds grotere complexiteit en de corresponderende dataverzamelingen van steeds grotere omvang. Wiskundige modellering van biologische processen, statistische analyse van biologische datasets, datamining en ontwikkeling van efficiënte computeralgoritmen kunnen voor de voortgang in kennis van de biologie veel betekenis hebben. Het nut voor de biologie van wiskundig modelleren en statistische analyse is uiteraard niet beperkt tot de thans zo populaire genomics. Ook is het gebruik van wiskunde in de biologie niet nieuw. Zeer bekend zijn de in de jaren 1930–'40 door Lotka en Volterra ontwikkelde modellen voor populatiegroei die gebaseerd zijn op differentiaalvergelijkingen, en het statistische werk van Fisher in de genetica van rond dezelfde tijd. Anderzijds hebben deze wiskundige toepassingen in de biologie de ontwikkeling van nieuwe wiskundige theorieën gestimuleerd.

In dit artikel geven we een tweetal voorbeelden uit eigen ervaring van biologische problemen waar wiskunde geholpen heeft de oplossing een stapje dichterbij te brengen. We willen hiermee niet alleen illustreren dat wiskunde een nuttige rol kan spelen bij het oplossen

van heel verschillende biologische problemen, maar ook dat het bij het gebruik van de wiskunde om meer gaat dan het oplossen van een simpele differentiaalvergelijking of het uitrekenen van een p -waarde. Elk biologisch probleem heeft een eigen benadering nodig. Zo zijn de benodigde wiskundige technieken zeer uiteenlopend van aard. Bovendien is soms geen van de bestaande technieken geschikt en dient er nieuw wiskundig gereedschap ontwikkeld te worden. In onderstaande voorbeelden gaat het in het bijzonder om stochastische modellering en statistische analyse van twee biologische processen.

De voorbeelden betreffen het ontstaan van kanker en ionkanalenkinetiek, twee zeer complexe biologische processen. In zekere zin bouwen beide voorbeelden voort op eerder werk van de auteur op het gebied van de modellering en analyse van de groei van plantencelpopulaties in batchcultures (De Gunst (1989)). In het eerste voorbeeld gaat het namelijk ook om het modelleren van de groei van celpopulaties. Het groeimodel voor het totaal aantal cellen in de populatie is van een eenvoudiger type dan dat voor de plantencellen in een batchculture, maar een extra dimensie die de situatie compliceert, is dat de ruimtelijke rangschikking van de cellen van belang is en dat er meerdere populaties tegelijkertijd kunnen bestaan. Het tweede voorbeeld gaat niet over het modelleren van celpopulaties, maar wel over plantencellen. De ionkanalen die in dit voorbeeld worden bestudeerd, zijn kaliumkanalen in protoplasten van bladcellen van gerst.

Gemeenschappelijk hebben de twee voorbeelden het volgende schema. Centraal staat een biologische vraagstelling. Om deze te beantwoorden zijn experimenten gedaan, waarmee gegevens (data) zijn verzameld. Voor statistische analyses van de data is het nodig een wiskundig model voor de data te hebben. Dit model wordt verkregen door eerst het biologisch proces waarin men is geïnteresseerd wiskundig te modelleren en van daaruit een model voor de data af te leiden. Het wiskundige model voor het biologische proces gebruikt zoveel mogelijk bestaande biologische kennis. Voorts worden de data statistisch geanalyseerd op basis van het datamodel. Tot slot worden op grond van de analyse conclusies getrokken. Waar nodig kunnen vervolgens nieuwe biologische experimenten worden gedaan en kan het zelfde schema worden herhaald. Dit schema is niet voorbehouden aan deze specifieke voorbeelden, maar beschrijft in het algemeen de volgorde waarin een gedegen statistische analyse zou moeten worden uitgevoerd.

biologische processen

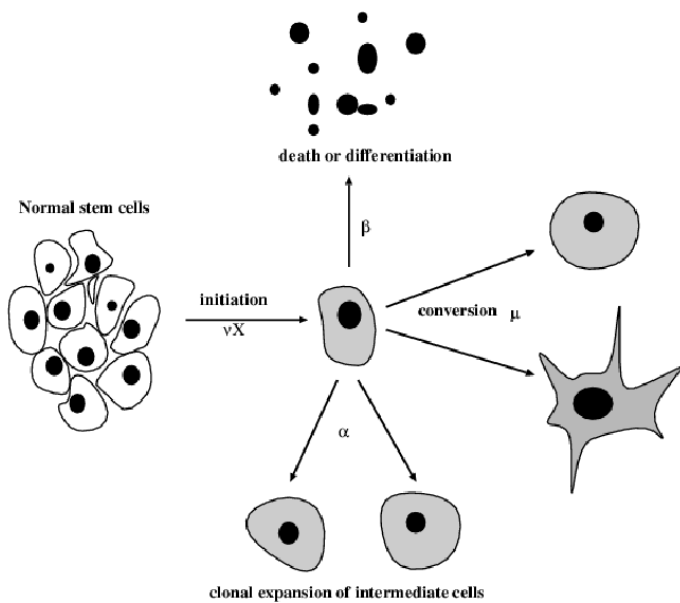
Natuurlijk zal geen van de gekozen wiskundige modellen de biologische werkelijkheid exact beschrijven. Deze werkelijkheid is immers veel te complex. Het is essentieel dat een model de werkelijkheid zodanig vereenvoudigt dat er nog goed met het model te rekenen valt, terwijl de details die van belang zijn, in het model zijn opgenomen. Omdat in de voorbeelden sprake is van verschijnselen die een zekere willekeur vertonen, zijn de gebruikte modellen stochastische modellen. We zullen enige wiskundige terminologie gebruiken. Voor basisbegrippen van de kansrekening en statistiek verwijzen we naar, bijvoorbeeld, Feller (1991), respectievelijk, Rice (1995). Voor een introductie in de stochastische processen is Karlin en Taylor (1997) te raadplegen. Veelvuldig komt in het vervolg het woord 'Markov' voor. Een proces is een Markovproces of heeft de Markoveigenschap wanneer we genoeg hebben aan kennis omtrent de huidige toestand van het proces (het heden) om de kans op toekomstige toestanden van het proces te kunnen bepalen. Het verleden doet dus niet ter zake. Dit in tegenstelling tot, bijvoorbeeld, een proces in de tijd waarbij wat de afgelopen maanden gebeurde, van invloed is op de toekomstige stand van zaken. Dergelijke processen kunnen niet met een Markovproces worden gemodelleerd. Het welbekende Poissonproces is een voorbeeld van een Markovproces.

In de voorbeelden laten we ook enkele in de afgelopen jaren ontwikkelde en populair geworden wiskundige methoden en concepten de revue passeren. Zo komt de Bayesiaanse statistiek aan de orde (zie Bernardo en Smith (1994) voor een introductie). Deze tak van de statistiek, die gebruik maakt van voorkennis omtrent de parameterwaarden, is genoemd naar Thomas Bayes, die in 1763 het als 'Bayes' Theorem' bekende resultaat voor voorwaardelijke kansen presenteerde. Een resultaat dat veel wordt gebruikt in de Bayesiaanse statistiek. Tot voor kort werd de Bayesiaanse statistiek volledig gescheiden van de klassieke, frequentistische, statistiek beoefend en had zij voornamelijk in de economie haar toepassing. Door onderzoek naar de theoretische achtergronden komen beide takken van de statistiek echter steeds dichterbij elkaar en Bayesiaanse statistiek wordt meer en meer toegepast, met name in de medische en biologische wetenschappen. Bij het populairder worden van de Bayesiaanse statistiek hebben de zogenaamde Markov chain Monte Carlo (MCMC)-methoden een katalyserende rol gespeeld. Deze methoden zijn geïntroduceerd door Me-

tropolis et al. (1953) en Hastings (1970). In die tijd werden ze vooral gebruikt voor het simuleren van systemen in de statistische fysica. Vanaf 1990 werd het gebruik van MCMC in de statistiek gemeengoed door de introductie van de zogenaamde Gibbs sampler als middel om steekproeven uit een hoger dimensionale verdeling te simuleren (Gelfand en Smith (1990)). Hoewel ze niet gebonden zijn aan een Bayesiaanse context, kunnen MCMC-methoden, zoals hieronder uitgelegd, in sommige situaties waar een klassieke benadering problemen oplevert, uitkomst bieden. Voor meer informatie over MCMC verwijzen we naar Gilks et al. (1996). Een belangrijk modern concept dat aan de orde komt, is het 'hidden Markovmodel'. Een hidden Markovmodel is gedefinieerd als een, eventueel stochastische, functie van een Markovproces. Statistische analysemethoden voor hidden Markovmodellen zijn reeds geïntroduceerd door Baum en Petrie (1966). Pas sinds eind jaren 1980 is de toepassing van deze modellen echt ontdekt, als eerste in de spraakherkenning. Momenteel zijn de modellen heel populair en worden ze niet alleen voor spraakherkenning gebruikt, maar ook voor economische en industriële toepassingen, voor weersvoorspellingen en in de biologie. Biologische toepassingen waar hidden Markovmodellen worden gebruikt, zijn, onder andere, biologische sequentieanalyse, genetische koppelinganalyse en fylogenetische boomreconstructie. Een heldere introductie tot hidden Markovmodellen wordt gegeven in Rabiner en Juang (1986).

Chemische carcinogenese

Dit voorbeeld betreft observatie en studie van vroege carcinogenese. De vraagstelling is of een bepaalde chemische stof kankerverwekkend is en, zo ja, waar in het groeiproces en hoe sterk de stof ingrijpt. In het bijzonder beschouwen we hier stoffen die gedacht worden tumorgroei in de lever te induceren. Daarom werden experimenten gedaan waarbij muizen of andere knaagdieren een chemische stof kregen toegediend gedurende verschillende perioden en in verschillende doseringen. Op verschillende tijdstippen na toediening werden de muizen gedood en werd hun lever onderzocht op bepaalde enzymdeficiënte celclusters. Deze enzymdeficiëntie is, naar men denkt, een voorstadium van kanker. Gedurende de tijd van de experimenten die we hier beschouwen, zullen kwaadaardige transformaties nog niet optreden. Voor situaties waarin dit wel het geval is, verwijzen we naar De Gunst en



Figuur 1 Het twee-gebeurtenissenmodel van Knudson, Moolgavkar en Venzon. Voordat een cel kwaadaardig is, dienen er tenminste twee onomkeerbare veranderingen op te treden.

Luebeck (1994). Aantal en grootte van de enzymdeficiënte clusters in een of meer tweedimensionale doorsneden van kleine stukjes weefsel werden genoteerd. De oorspronkelijke vraag kan nu vertaald worden naar: wat is de invloed van de hoeveelheid toegediende stof op het aantal en op de grootte van enzymdeficiënte celclusters?

Procesmodel

Om tot een model voor de data te komen dient eerst een model voor het ontstaan en de groei van de enzymdeficiënte celclusters geformuleerd te worden. Aangezien dit proces gedacht wordt een onderdeel te zijn van de carcinogenese, kunnen we voor de keuze van een model een onderdeel van een model voor kankergroei nemen. Een uitgebreid overzicht van verschillende modellen voor kankergroei is te vinden in Kopp-Schneider (1997) of in Van Leeuwen en Zonneveld (2001). We kiezen voor het twee-gebeurtenissenmodel (two-event model), ook wel MVK-model genoemd naar degenen die dit model voor het eerst in de context van kankermodellering gebruikten (Knudson (1971); Moolgavkar en Venzon (1979)). Het onderliggende idee van dit model is dat, voordat een cel kwaadaardig is, er tenminste twee onomkeerbare veranderingen, bijvoorbeeld mutaties, moeten zijn opgetreden. We nemen aan dat de eerste verandering leidt tot een cel met enzymdeficiëntie. We beschrijven nu eerst het gedeelte van het MVK-model dat wij nodig hebben (zie figuur 1).

Zij X het aantal gezonde cellen per volume-eenheid in de lever. Deze cellen lopen het risico een verandering te ondergaan die zich uit in enzymdeficiëntie. De verandering wordt aangenomen onomkeerbaar te zijn. De kans dat de verandering optreedt is heel klein. Wanneer een cel de verandering heeft ondergaan, dan kan zij bij deling de verandering doorgeven aan haar nakomelingen. Er ontstaat dan een enzymdeficiënt celcluster, dat vanwege de gemeenschappelijke voorouder — dit is de cel die als eerste veranderde — een enzymdeficiënte celkloon wordt genoemd.

Zij ν de snelheid per tijdseenheid (intensiteit) waarmee normale cellen de verandering kunnen ondergaan. Dan ontstaan in een volume-eenheid de enzymdeficiënte, pre-kwaadaardige, celklonen volgens een Poissonproces met intensiteit νX . Het aantal enzymdeficiënte celklonen per volume-eenheid, dat we N_V zullen noemen, is dus een sto-

chastisch aantal. Dit in tegenstelling tot het aantal gezonde cellen per volume-eenheid, X . Dat aantal wordt vast genomen omdat het zoveel groter is dan het aantal N_V waarin we zijn geïnteresseerd, dat de fluctuaties erin verwaarloosbaar klein zijn. Indien een enzymdeficiënte kloon niet zou kunnen uitsterven, zou als we op tijd 0 met alleen gezonde cellen begonnen, na een tijd t het aantal N_V Poisson verdeeld zijn met verwachting $\nu X t$. Echter, een enzymdeficiënte celkloon kan wél uitsterven, namelijk wanneer in de kloon er telkens meer cellen doodgaan of nog een tweede verandering ondergaat zodat ze kwaadaardig zijn geworden, dan er door celdeling bijkomen. Celdeling en celsterfte modelleren we met een geboorte-sterfteproces (zie, bijvoorbeeld, Karlin en Taylor (1997)) waarin de enzymdeficiënte cellen in een kloon, onafhankelijk van elkaar, zich delen met intensiteit α en doodgaan met intensiteit β . Het aantal enzymdeficiënte celklonen per volume-eenheid na een tijd t is nog steeds Poisson verdeeld, maar nu met een verwachting, E_V , die een functie is van νX , t , α en β . We hebben

$$E_V = -\frac{\nu X}{\alpha} \log(1 - \alpha/\beta p(t)),$$

waarbij $p(t)$ wordt gegeven door

$$p(t) = \frac{1 - \exp(-(\alpha - \beta)t)}{\alpha/\beta - \exp(-(\alpha - \beta)t)}.$$

Wanneer we ook nog aannemen dat de verschillende klonen zich onafhankelijk van elkaar ontwikkelen, dan kunnen we de kansverdeling van de grootte van de klonen bepalen. We symboliseren deze verdeling door p_1, p_2, \dots, p_M , waarbij p_m de kans is dat een kloon uit m cellen bestaat en M een willekeurig getal dat groter is dan het grootste aantal cellen waaruit een kloon aan het eind van het experiment redelijkerwijs kan bestaan. Op tijd t is p_m een functie van m , t , α en β , maar niet van νX :

$$p_m = -\frac{(\alpha/\beta p(t))^m}{m \log(1 - \alpha/\beta p(t))}$$

met $p(t)$ als boven. Het model voor ontstaan en groei van enzymdeficiënte celklonen is nu volledig.

Als de chemische stof vooral het ontstaan van de klonen zal versnellen, dan zal de dosis van de stof invloed hebben op de waarde van ν ; de stof heet dan een initiator. Als de chemische stof vooral de groei van de klonen zal versnellen, zal de dosis van de stof invloed hebben op de waarde van $\alpha - \beta$; de stof heet dan een promotor. De waarden van de parameters ν , α en β zijn echter onbekend en zullen geschat dienen te worden op grond van de data. Wanneer onze data hadden bestaan uit aantal en grootte van de celklonen zelf, dan had dit eenvoudigweg kunnen gebeuren met behulp van de meest aannemelijke (maximum likelihood) schatters. De aannemelijkheidsfunctie die bij het model hoort, dat is de kans op de data onder het model gezien als functie van de onbekende parameters, wordt dan bepaald en gemaximaliseerd naar die parameters. Echter, de data bestaan uit aantal en groottes van doorsnijdingen van de celklonen, ook wel transecties of foci genoemd, geobserveerd in een of meer tweedimensionale weefseldoorsneden. Dit betekent dat het hierboven geformuleerde model nog geen model is voor de data. Immers, het model beschrijft aantal en groottes van driedimensionale objecten, terwijl de data aantal en groottes van tweedimensionale objecten betreffen. Deze tweedimensionale objecten hebben geen één-op-één-relatie met de driedimensio-

nale objecten, want een kleine transectie kan zowel afkomstig zijn van een grote als van een kleine kloon. Ook zullen transecties van grotere klonen meer kans hebben op een weefseldoorsnede voor te komen dan die van kleinere klonen.

Datamodel

Het schatten van driedimensionale grootheden op grond van tweedimensionale data is een vaak voorkomend, meestal lastig statistisch probleem, behorend tot de klasse van inverse problemen. Het hier geschetste stereologische probleem van het schatten van aantal- en grootteverdeling van objecten op grond van observaties aan een vlakke doorsnede is uitgebreid bestudeerd sinds Wicksell (1925) de oplossing presenteerde voor bolvormige objecten (hier: de enzymdeficiënte celklonen). Een overzicht van verschillende methoden om dit probleem aan te pakken, is te vinden in Stoyan, Kendall en Mecke (1995). De meeste methoden zijn niet-parametrisch en betreffen aselechte steekproeven en bolvormige objecten. Meer recent zijn ook methoden ontwikkeld voor systematische steekproeven, en andere dan bolvormige objecten (Gundersen et al. (1988)). In sommige situaties is een parametrische aanpak mogelijk.

In onze experimenten zijn de weefseldoorsneden willekeurig gekozen en hebben we dus te maken met aselechte steekproeven. Bovendien ligt, omdat we een parametrisch model voor de klonale groei hebben, een parametrische methode voor de hand. In sommige experimenten zien de transecties er min of meer cirkelvormig uit. Dan kunnen de klonen bolvormig worden verondersteld. In de meeste studies zijn de transecties van allerlei vorm en is de aanname van bolvormige klonen onterecht. Er is geen standaard parametrische methode voorhanden voor aselechte steekproeven. We moeten deze dus zelf ontwikkelen voor onze situatie.

Wat we willen is het driedimensionale model voor klonale groei vertalen naar een tweedimensionaal model voor de data dat dezelfde onbekende parameters heeft als het driedimensionale model. Het tweedimensionale model zal bestaan uit de kansverdeling van het aantal en die van de groottes van de transecties. Om deze kansverdelingen te bepalen hebben we niet genoeg aan het driedimensionale groeimodel, maar dienen we ook aannamen te doen over de ruimtelijke structuur van de celklonen, met name over waar ze zich bevinden en hoe ze groeien in de ruimte.

Wat het eerste betreft nemen we aan dat de cellen min of meer bolvorming en even groot zijn en dat de middelpunten van de cellen die als eerste de verandering ondergaan, een Poissonproces in de ruimte vormen. Wat het tweede betreft, kunnen we ons verschillende structuren voorstellen. Bijvoorbeeld, een waarin een nieuwe cel ontstaat uit deling op een willekeurige plek, of juist op een zodanige plek dat in de celclusters die ontstaan, de cellen zo dicht mogelijk opeengepakt zitten. In eerste instantie kunnen we de celklonen ook als bollen beschouwen (Moolgavkar et al. (1990)), maar vooral bij kleinere klonen is deze benadering te grof. In De Gunst en Luebeck (1998) worden twee verschillende modellen doorgerekend waarin geen bolvorm voor de klonen wordt verondersteld (zie figuur 2).

Welke ruimtelijke configuratie we voor de klonen ook kiezen, essentieel is dat de klonen groeien onafhankelijk van de positie van hun eerste cel en onafhankelijk van andere klonen, en dat het groeiproces isotroop is, dat wil zeggen, geen voorkeursrichting heeft. Dan blijkt het namelijk mogelijk om voor elk tijdstip de kansverdeling van het aantal waargenomen transecties per eenheidsoppervlak, dat we N_O zullen noemen, en die van de grootte van de transecties, gesymboliseerd door de kansen q_1, q_2, \dots, q_M , te bepalen. N_O is weer Poisson

verdeeld met verwachting E_O , die afhangt van de verwachting E_V van N_V , van de p_m en van termen α_{nm} ($n = 1, \dots, m$, $m = 1, \dots, M$) die afkomstig zijn van de ruimtelijke structuur van de klonen. In formule:

$$E_O = E_V \sum_{m=1}^M p_m \sum_{n=1}^m \alpha_{nm}.$$

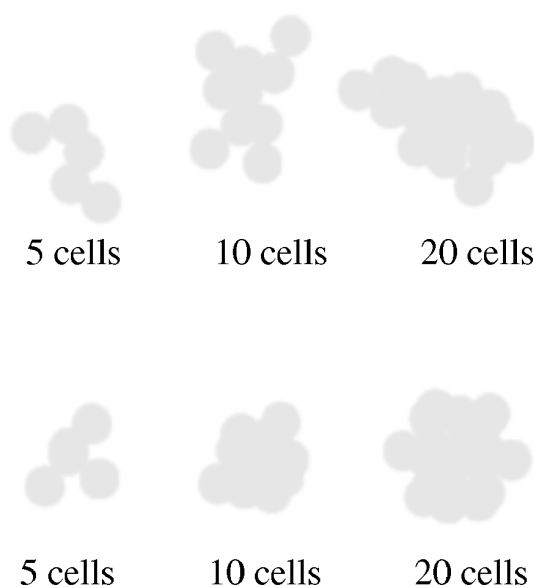
De kans q_n dat een transectie op een bepaald tijdstip uit n cellen bestaat blijkt uitgedrukt te kunnen worden in termen van p_m en de α_{nm} . We hebben

$$q_n = \frac{\sum_{m=n}^M p_m \alpha_{nm}}{\sum_{m=1}^M p_m \sum_{n=1}^m \alpha_{nm}}.$$

Dit maakt het tweedimensionale wiskundige model voor de data compleet.

De kansverdelingen van N_O en van de grootte van de transecties hangen dus af van de tijd en van de onbekende parameters ν , α en β . Verder hangen ze alleen via de termen α_{nm} af van de ruimtelijke structuur. De coëfficiënt α_{nm} is de kans dat in een eenheidsvolume een willekeurige kloon ter grootte m doorsneden wordt en dat de transectie n cellen laat zien. De termen α_{nm} zijn onafhankelijk van het klonale groeimodel en van de tijd en daardoor kunnen ze bepaald worden. Voor heel simpele ruimtelijke configuraties kunnen ze expliciet worden berekend; voor biologisch plausible configuraties, die doorgaans complex van aard zijn, kan dit meestal niet. Wel is het dan mogelijk de termen te benaderen met behulp van computersimulaties van de ruimtelijke configuraties.

Door willekeurige doorsnijdingen aan te brengen kan de fractie van de klonen met m cellen die doorsneden wordt en daarbij n cellen in de transectie laat zien, geteld worden. Bij een groot genoeg aantal gesimuleerde configuraties is deze fractie een goede benadering van de α_{nm} . Deze procedure is alleen mogelijk voor niet al te grote m en n . Komen er ook grote transecties in het experiment voor, dan kunnen we zogenaamde cellulaire automaten (Toffoli en Margolus, 1989) gebruiken om de waarden van α_{nm} te benaderen (Luebeck en De Gunst (2001)).



Figuur 2 Gesimuleerde clusters volgens twee ruimtelijke modellen. Boven: deling willekeurig. Onder: opeengepakt. (Illustratie uit De Gunst en Luebeck (1998).)

Statistische analyse

Nu we een manier hebben om de α_{nm} te berekenen, zijn de enige onbekende grootheden in de kansverdelingen van N_O en van de transectiegroottes, en dus in die van de data, de parameters ν , α en β . Dit zijn precies de parameters die we wilden schatten. Met de data en hun kansverdelingen volgens het tweedimensionale datamodel bepalen we nu de aannemelijkheidsfunctie en maximaliseren deze naar de parameters. Hiermee verkrijgen we de meest aannemelijke schattingen voor ν , α en β . Expliciete uitdrukkingen voor de schatters hebben we niet. We verkrijgen hun waarden door numerieke benadering. Om een indruk te krijgen van de invloed die dosering op de verschillende parameters heeft, moeten we een en ander natuurlijk doen voor de verschillende doses van de chemische stof. Een alternatief voor het schatten van de parameters voor elke dosering apart, is om tevens de afhankelijkheid van de dosis te modelleren en vervolgens de afhankelijkheidsparameters te schatten. Ook kan het effect van een stof onder verschillende regimes van toediening (een enkele injectie vs. continue toediening) onderzocht worden.

Samengevat gebruikt boven beschreven analysemethode alle beschikbare kwantitatieve informatie betreffende aantal en grootte van de transecties en resulteert zij in schattingen van biologisch relevante grootheden als aantal pre- kwaadaardige cellen per tijdseenheid en de delingssnelheid. Voorbeelden van analyses van experimentele data met de beschreven methode zijn, onder andere, te vinden in Moolgavkar et al. (1990), Luebeck et al. (1991), De Gunst en Luebeck (1998) en Grasl-Kraupp et al. (2000). In de verschillende analyses is, bijvoorbeeld, gevonden dat de stof N-nitrosomorfoline (NNM) een sterke initiator en een zwakke promotor is, is de invloed van deze stof op apoptose (geprogrammeerde celsterfte) onderzocht en zijn de promotiesterktes van verschillende polychlorinated bifenylen (PCBs) met elkaar vergeleken. Ter illustratie zien we in figuur 3 experimentele data met het gefitte model, zowel voor aantal als voor grootte van de transecties.

Heterogeniteit, de Bayesiaanse methode en MCMC

Tot slot bespreken we nog een ander lastig punt: heterogeniteit. Waar bovenstaand model geen rekening mee houdt, is dat de proefdieren, hoewel zoveel mogelijk gelijkwaardig gekozen, niet noodzakelijk hetzelfde op een zelfde behandeling zullen reageren. Bij grote verschillen tussen de dieren is een model met dezelfde groeiparameters voor alle dieren niet meer van toepassing. Geven we elk dier een set eigen parameters, dan wordt het aantal te schatten parameters erg groot en de interpretatie van de resultaten lastig. Een tussenweg is om ook de

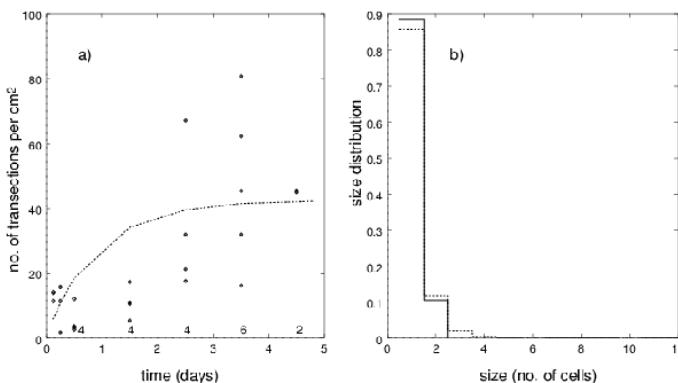
parameters die voor elk dier een andere waarde lijken te hebben, stochastisch te laten zijn en te veronderstellen dat de parameterwaarden van elk dier uit een gezamenlijke onderliggende kansverdeling met enkele nieuwe, onbekende, parameters afkomstig zijn. We zijn er dan niet meer op uit om de parameterwaarden van elk van de dieren te schatten, maar die van de onderliggende verdeling. Doorgaans zijn dat er veel minder. Laten we de parameterwaarden van de onderliggende kansverdeling hyperparameters noemen. De aannemelijkheidsfunctie voor deze hyperparameters wordt in onze situatie helaas nogal complex. Dit vanwege het feit dat we de parameters van elk proefdier uit integreren en er dus evenzoveel integralen in de te maximaliseren aannemelijkheidsfunctie komen te staan. Het schatten van de hyperparameters met behulp van meest aannemelijke schatters, zal daarom praktisch niet meer doenlijk zijn. Een goed alternatief is een Bayesiaanse aanpak waarbij de a-posterioriverdeling gegenereerd wordt met behulp van een MCMC-methode.

De essentie van een Bayesiaanse aanpak is, dat we voordat we een experiment doen vaak al een idee hebben over de parameterwaarden. De parameters kunnen, bijvoorbeeld, alleen maar positief zijn of moeten binnen bepaalde grenzen liggen. Deze ideeën over de parameters kunnen we samenvatten door voor de parameters een geschikte kansverdeling te vooronderstellen. Deze kansverdeling wordt de a-prioriverdeling genoemd. Nadat het experiment is uitgevoerd, leveren de data van het experiment ons nieuwe informatie over de verdeling van de parameters, zodat we de a-prioriverdeling moeten bijstellen. De aldus verkregen bijgestelde verdeling is de zogenaamde a-posterioriverdeling. De a-posterioriverdeling en de a-prioriverdeling hangen met elkaar samen via de aannemelijkheidsfunctie:

$$\pi(\theta|y) = \frac{l(y|\theta)p(\theta)}{\int l(y|\theta)p(\theta)d\theta} \propto l(y|\theta)p(\theta).$$

Hierin representeren y de data, θ de parameters, $\pi(\theta|y)$ de a-posterioriverdeling, $p(\theta)$ de a-prioriverdeling en $l(y|\theta)$ de aannemelijkheidsfunctie. Het symbool \propto betekent 'is evenredig met'; de noemer in de tweede term hangt niet van θ af en is dus een constante. Meestal is het niet mogelijk deze constante expliciet te berekenen, waarmee de a-posterioriverdeling bekend is op een evenredigheidsconstante na. Immers, de aannemelijkheidsfunctie volgt uit het model en de a-prioriverdeling kiezen we zelf, dus deze componenten van de a-posterioriverdeling zijn bekend.

In de Bayesiaanse statistiek wil men niet de parameterwaarden, maar de a-posteriorikansverdeling van de parameters schatten. (We merken op dat onder bepaalde voorwaarden en bij voldoende waarnemingen de modus van de a-posterioriverdeling praktisch gelijk is aan de waarde van de meest aannemelijke schatter.) Om een verdeling te schatten zouden we willen beschikken over een steekproef uit die verdeling. Dan is, bijvoorbeeld, een histogram van de steekproef een schatting van de kansverdeling. Aangezien we de verdeling niet kennen, kunnen we er niet zo maar een steekproef uit genereren. Met MCMC-methoden kan dit wel, mits de verdeling bekend is op een evenredigheidsconstante na. Dergelijke methoden genereren namelijk realisaties van een Markovketen die als stationaire verdeling precies de onbekende verdeling heeft waarin we zijn geïnteresseerd. Bij de eerste generatiestappen is de keten nog niet stationair, maar na een tijdje (de burn-in-periode) wel. De realisaties die na de burn-in-periode gegenereerd worden, vormen dan een steekproef uit de onbekende verdeling. Algoritmen voor MCMC zijn doorgaans eenvoudig op te stellen. Het principe is als volgt. Er wordt begonnen met, eventueel wille-



Figuur 3 a. Waargenomen (open cirkels) en geschatte (stippellijn) aantal transecties per vierkante centimeter. Bij de tijds staat het aantal dieren dat op dat tijdstip is gebruikt. Een waarneming, 165 op tijdstip 3,5, ligt buiten het afgebeelde gedeelte van de grafiek. b. Waargenomen grootteverdeling (ononderbroken lijn) en geschatte grootteverdeling van de transecties. (Illustratie uit De Gunst en Luebeck (1998).)

keurige maar liefst goed gekozen, aanvangswaarden voor de parameters. Vervolgens wordt telkens een volgende realisatie verkregen door kandidaat-waarden voor de parameters te genereren uit een handig gekozen verdeling (de proposalverdeling). De kandidaat-waarden worden al dan niet geaccepteerd op grond van vergelijking van nieuwe en oude waarden volgens een criterium gebaseerd op de a-prioriverdeling en de aannemelijkheidsfunctie. Bij acceptatie vormen de nieuwe waarden de nieuwe realisatie; bij niet-acceptatie blijven de oude waarden staan. Er dienen na de burn-in-periode voldoende realisaties gegenereerd te worden opdat de steekproef representatief is. Toepassing van MCMC-methoden is niet altijd zonder problemen. Goede aanvangswaarden dienen gevonden te worden, goede keuzes dienen te worden gemaakt voor de a-prioriverdeling, het bepalen van de lengte van de burn-in-periode is vaak lastig en het opzetten van een efficiënt algoritme en de interpretatie van de resultaten vereisen de nodige ervaring. Figuur 4 laat twee typische voorbeelden van MCMC zien. Vooral in de tweede reeks is duidelijk te zien dat de burn-in-periode daar ongeveer 100 generatiecycli lang is.

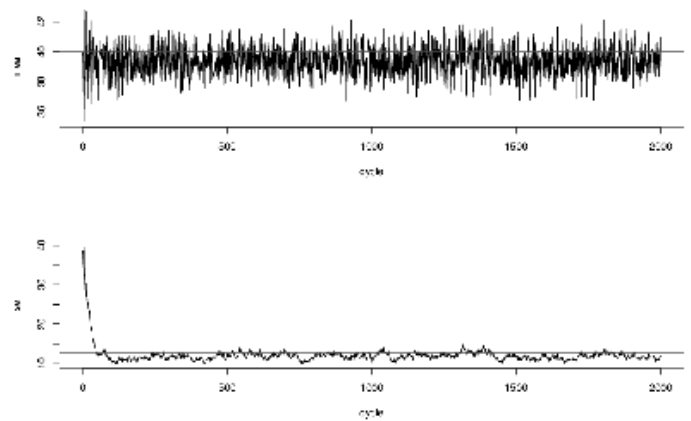
In De Gunst et al. (2003) is heterogeniteit tussen de proefdieren van de enzymdeficiëntie experimenten onderzocht met behulp van MCMC-methoden. Inderdaad bleek het opnemen van heterogeniteit in het model voor de onderzochte data een verbetering ten opzichte van een model zonder heterogeniteit. Verrassend was, dat niet voor alle parameters een aanname van heterogeniteit nodig is. Voor sommige parameters voldoet de aanname van homogeniteit prima.

Ionkanalenkinetiek

Ionkanalen zijn complexe eiwitten in de celmembraan die in bepaalde configuraties tunnelvormig zijn, zodat er ionen doorheen kunnen die zo het membraan kunnen passeren (Hille (1992)). Ze spelen niet alleen een cruciale rol bij het in stand houden van de energetische en osmotische balans in levende cellen, maar zijn ook essentieel voor inter- en intracellulaire communicatie in cellulaire signaalprocessen. De regulatie en de functie van het openen en sluiten van de kanalen worden uitgebreid bestudeerd in de biofysica, biochemie en fysiologie, maar zijn nog steeds niet volledig begrepen. Wanneer er ionen door een ionkanaal gaan, resulteert dit in een elektrische stroom die gemeten kan worden met behulp van de zogenaamde patchclamptechniek (Hille (1992); Sackman en Neher (1995)). Het idee is, door wiskundige modellering van het proces van openen en sluiten (het 'gating'-mechanisme) en daarna analyse van gemeten stromen met behulp van dit model, meer te weten te komen over de regulatie van de kanalen.

Procesmodel

Afhankelijk van het type meting kan de stroom afkomstig zijn van een of van een paar kanalen of van alle kanalen in een hele cel. Wij hebben ons in eerste instantie beziggehouden met het modelleren en analyseren van metingen aan een enkel kanaal. Figuren 5a en 5c laten een gedeelte van zo'n meting zien. Een kanaal kan zich in verschillende configuraties bevinden, waarvan sommige (de open toestanden) een ionenstroom mogelijk maken en andere (de gesloten toestanden) niet. De verandering in configuratie over de tijd vertoont een zekere willekeur. In het algemeen worden Markovketens gebruikt om deze willekeur te beschrijven (Colquhoun en Hawkes (1995)). Een Markovketen is een Markovproces met een eindig of aftelbaar aantal toestanden. Hier gaat het om een eindig aantal. De verschillende configuraties waarin het ionkanaal zich kan bevinden, zijn de toestanden van de Markovketen en het kanaal gaat van de ene toestand over in een andere toestand volgens de regels van de Markovketen (zie figuur 6 voor een



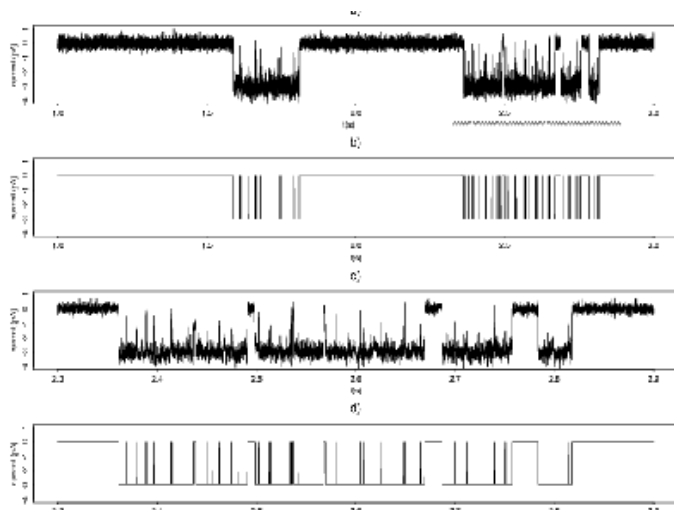
Figuur 4 Twee MCMC-reeksen

paar voorbeelden van mogelijke Markovketens). Het probleem wordt dan om op grond van de data — de gemeten stromen — het juiste model te vinden en de daarbij behorende overgangskansen tussen de verschillende toestanden te schatten. Dit zou op simpele wijze gedaan kunnen worden als we wisten hoeveel toestanden er bestaan en van welke soort deze zijn, als we wisten van welke toestand naar welke toestand het kanaal kan overgaan, en als we konden waarnemen in welke achtereenvolgende toestanden het kanaal zich gedurende een meetexperiment bevindt. Helaas is aan geen van deze voorwaarden voldaan. Hoewel in het algemeen wordt aangenomen, dat er meer dan één open of gesloten toestand per kanaal is, is niet bekend hoeveel dit er zijn en tussen welke toestanden overgangen mogelijk zijn. Bovendien wordt in de meeste experimenten slechts waargenomen of een kanaal open is of niet, maar kan er geen onderscheid worden gemaakt tussen verschillende open, respectievelijk, gesloten toestanden. Dit betekent dat we, net als in de vorige sectie, te maken hebben met een lastig inverse probleem. Dit probleem wordt nog verder gecompliceerd doordat de metingen worden verstoord door ruis, de opnames van de stroom gesampled zijn en er filtereffecten optreden, waardoor zeer korte openingen en sluitingen niet of niet goed kunnen worden gedetecteerd. Kortom, ook nu is het model dat we hebben geformuleerd voor het proces dat ons interesseert, dit is de configuratieverandering van het kanaal, niet een model voor de data.

Datamodel

Allereerst maken we ons even niet druk om welke Markovketen geschikt is om het gating-mechanisme te beschrijven. Vervolgens merken we op dat de data, dat wil zeggen het proces dat we waarnemen, de waargenomen stroom, eigenlijk een functie is van deze Markovketen. Immers, de verschillende open, respectievelijk, gesloten toestanden kunnen niet van elkaar worden onderscheiden. We zien in onze opnamen slechts wel of niet stroom lopen. Bovendien zorgt de meetapparatuur ervoor dat over het geheel ruis en een filter worden gelegd. Dit betekent, dat we de data zouden kunnen beschrijven met een hidden Markovmodel. Er zijn hoofdzakelijk drie problemen die door middel van een statistische analyse op basis van hidden Markovmodellen kunnen worden opgelost en die alle drie voor ons van belang zijn:

- gegeven de data en het model inclusief parameterwaarden, hoe de kans op de waargenomen data te berekenen;
- gegeven de data en het model inclusief parameterwaarden, hoe de meest waarschijnlijke onderliggende realisatie van de Markovketen (in ons geval de werkelijke onderliggende reeks configuraties



Figuur 5 a) Een gedeelte van een opname (2.0 sec) aan een kalium outward-rectifier in een protoplast van een bladcel van gerst. b) Het met MCMC geschatte onderliggende open-geslotenproces behorend bij de opname in a). c) Het gedeelte van de opname dat gemarkeerd is in a). d) Het met MCMC geschatte onderliggende open-geslotenproces behorend bij de opname in c). (Illustratie uit De Gunst et al. (2001).)

tijdens de opname van de data) te bepalen;

- iii. gegeven de data en het model, hoe de parameterwaarden te bepalen opdat de kans op de waargenomen data maximaal is.

De standaard algoritmen voor het oplossen van deze problemen heten respectievelijk de forward-backward procedure, het Viterbi-algoritme en de Baum-Welch herschattingsmethode. Voor het modelleren van ionkanalenkinetiek werden eerder hidden Markovmodellen gebruikt; voor het eerst in Chung et al. (1990). De in dat artikel en latere publicaties over ionkanalen gebruikte hidden Markovmodellen zijn echter voor onze data niet geschikt.

Onze data bestaan uit metingen aan de K^+ -outward-rectifier in protoplasten van bladcellen van gerst. Dit kanaal is een spanningsgereguleerd kanaal. Het vertoont een bijzonder soort flikkergedrag, wat maakt dat de bij het meten gebruikte filter waarschijnlijk veel invloed op de data heeft. Ook zien we dat de ruis groter is wanneer het kanaal open is, dan wanneer het gesloten is. Tenslotte blijkt de ruis niet wit te zijn. Onze onderzoeken hebben uitgewezen dat het opnemen van deze drie aspecten in het hidden Markovmodel voor de gemeten stroom essentieel is (Schouten, 2002). Eerder gebruikte hidden Markovmodellen hebben niet al deze componenten tezamen in het model.

Hoe ziet ons hidden Markovmodel eruit? Zij X_t de toestand van het kanaal op tijd t , dat wil zeggen dat het proces $\{X_t\}_{t \geq 0}$ de Markovketen is die het gating-mechanisme beschrijft. Laat verder Y_t de gemeten stroom op tijd t zijn. Dan hebben we als model dat Y_t de volgende functie is van het proces $\{X_t\}_{t \geq 0}$:

$$Y_t = \sum_{k=-r}^r \gamma_k \mu(X_{t-k}) + C_t + \sigma(X_t) \delta_t.$$

De functie μ is de stroomgrootte. Deze is gelijk aan μ_g als de toestand een gesloten toestand is (er loopt altijd wat stroom, ook als het kanaal dicht is, de lekstroom) en aan μ_o als de toestand een open toestand is. Er geldt $\mu_o > \mu_g$. De eerste term in het rechterlid van het model representeert de filter: de γ_k zijn de (bekende) filtercoëfficiënten en we zien dat de gemeten stroom op een bepaald tijdstip een gewogen gemiddelde is van de stroom op dat tijdstip en enkele buurtijdstippen. De tweede term representeert de niet-witte ruis en is een autoregressief

proces (zie, bijvoorbeeld, Brockwell en Davis (1996)) gegeven door

$$C_t = \sum_{i=1}^p \phi_i C_{t-i} + \epsilon_t$$

met de ϵ_t onafhankelijk en normaal verdeeld met verwachting 0 en variantie σ_ϵ^2 . Ruis van een paar tijdstippen terug bepaalt de ruis van het huidige moment. De laatste term in het hidden Markovmodel voor Y_t representeert de toestandsafhankelijke component van de ruis; de δ_t zijn onafhankelijk en standaard normaal verdeeld. De functie σ beschrijft de grootte van de ruis en is gelijk aan σ_g als de toestand een gesloten toestand is en aan σ_o als de toestand een open toestand is. We hebben $\sigma_o > \sigma_g$. Naast de onbekende overgangskansen van de Markovketen $\{X_t\}_{t \geq 0}$, zijn ook de beide μ 's, de drie σ 's en de ϕ_i 's onbekend.

Statistische analyse en selectie Markovketen

De hierboven genoemde algoritmen voor het berekenen van meest aannemelijke schattingen voor hidden Markovmodellen, forward-backward plus Baum-Welch, leveren in ons geval geen biologisch plausible schattingen (Schouten (2000)). De aannemelijkheidsfunctie is blijkbaar zo complex, dat het niet goed mogelijk is haar maxima numeriek te bepalen. Ook in deze toepassing zijn we daarom overgegaan op een Bayesiaanse aanpak en hebben we MCMC-methoden gebruikt om de a-posterioriverdelingen van de onbekende parameters te schatten (De Gunst et al. (2001)). De algoritmen van de gebruikte MCMC-methoden zijn nu echter een stuk ingewikkelder, omdat het hidden Markovmodel veel complexer is dan het geboorte-sterfteproces in het vorige voorbeeld. Bij het genereren van de realisaties van de Markovketen die de a-posterioriverdeling als stationaire verdeling heeft, worden nu namelijk in elke generatiestap niet alleen waarden van de parameters gegenereerd, maar ook van het ruisproces $\{C_t\}$ en het configuratieproces $\{X_t\}$ voor t in de meetperiode. We verkrijgen zo niet alleen inzicht in de parameterwaarden, maar ook in de verdeling van de toestand van het kanaal op elk tijdstip. Op grond van simulatiestudies blijkt deze aanpak heel goed te werken.

Willen we deze procedure toepassen op echte data, dan zullen we nu de Markovketen dienen te specificeren die het gating-mechanisme beschrijft, of eerst op de een of andere wijze tot een goede keuze moeten komen. Op grond van biologische overwegingen, hebben we weliswaar een aantal kandidaat-modellen (zie figuur 6), maar welk daarvan het meest geschikt is, weten we niet. We willen ons dan ook liever niet vastpinnen op één model. In de niet-Bayesiaanse context is modelselectie meestal gebaseerd op een criterium dat de aannemelijkheidsfunctie zodanig corrigeert dat grotere, meer complexe modellen minder snel worden gekozen, zoals bijvoorbeeld het welbekende Akaike-informatiecriterium (AIC) (zie Brockwell en Davis (1996)). Binnen het kader van de door ons gebruikte Bayesiaanse aanpak kunnen we gebruik maken van een ander type selectiecriterium, de zogenaamde Bayesfactor (Kass en Raftery (1995)). Hiervoor moeten we het model ook zien als een parameter waarvan we een a-priori-idee hebben. De Bayesfactor voor twee modellen waartussen gekozen dient te worden, is namelijk gedefinieerd als de verhouding tussen de ratio's van hun a-posteriori- en hun a-priorikansen. Het model met de grootste ratio komt het meest in aanmerking. Hoe groter het verschil, hoe meer aanleiding de data geven aan dit model de voorkeur te geven. In ons geval kunnen we de Bayesfactor voor twee kandidaat-modellen niet expliciet uitrekenen, omdat we wederom de a-posterioriverdeling slechts kennen op

een evenredigheidsconstante na.

Wanneer we het model zien als een parameter waarvan we a-priorikennis hebben, kunnen we in principe de a-posterioriverdeling van het model, en daarmee de Bayesfactor, bepalen met behulp van een MCMC-methode. We nemen dan ook aan dat er een a-priorikansverdeling is voor het model die aan elk van de kandidaat-modellen een bepaalde kans geeft. Vinden we het ene kandidaat-model waarschijnlijker dan het andere, dan krijgt het ene model een hogere kans toegekend dan het andere. Hebben we vooraf geen voorkeur, dan krijgen alle kandidaat-modellen dezelfde kans. Vervolgens gaan we op grond van de data de a-priorikansverdeling bijstellen en verkrijgen we een a-posteriorikansverdeling voor het model die bijgestelde kansen toekent aan de kandidaat-modellen. Het kandidaat-model met de hoogste a-posteriorikans (*maximum a posteriori probability* of MAP) komt dan op grond van de data het meest in aanmerking. De Bayesfactor weegt als het ware de a-posteriorikansen door middel van de a-priorikansen. Hebben we aanvankelijk geen voorkeur voor een bepaald model, dan zijn de a-priorikansen gelijk en is een keuze op grond van MAP identiek aan die op grond van de Bayesfactor.

Nu is het vinden van de a-posterioriverdeling van het model met behulp van MCMC niet zo recht toe recht aan als dat voor de a-posterioriverdeling van gewone parameters. Dit komt omdat het genereren van modellen alleen maar in connectie met bijbehorende parameterwaarden kan gebeuren. Deze dienen dan in dezelfde generatiecyclus meegegegeneerd te worden. Er zijn problemen wanneer de parameters voor twee verschillende modellen verschillend in aantal zijn of een heel andere betekenis hebben. Immers, in die gevallen is, wanneer het MCMC-algoritme een kandidaat-realisatie bestaande uit het ene model met bijbehorende parameterwaarden, moet vergelijken met de huidige realisatie bestaande uit een ander model met bijbehorende parameterwaarden (om al dan niet tot acceptatie van de kandidaat-realisatie over te gaan), niet helder wat met wat vergeleken moet worden. Green (1995) heeft hiervoor een oplossing bedacht in de vorm van de zogenaamde 'reversible jump' MCMC. Hierbij worden op een slimme manier extra parameters ingevoerd om het aantal parameters voor de verschillende modellen gelijk te maken. Hebben we dan eenmaal een model geselecteerd, dan bepalen we de a-posterioriverdelingen van de bijbehorende parameters met MCMC als boven uiteengezet.

In De Gunst en Schouten (2003) hebben we een reversible jump MCMC geïmplementeerd voor selectie van een set kandidaat-modellen voor het gating-mechanisme van het K⁺-kanaal. We hebben de procedure getest aan de hand van gesimuleerde data. Implementatie van reversible jump MCMC bleek in de praktijk niet eenvoudig te zijn. Bovendien eindigen af en toe twee modellen die veel op elkaar lijken, met nagenoeg dezelfde geschatte a-posteriorikans. Dit maakt een keuze op grond van deze kansen onmogelijk.

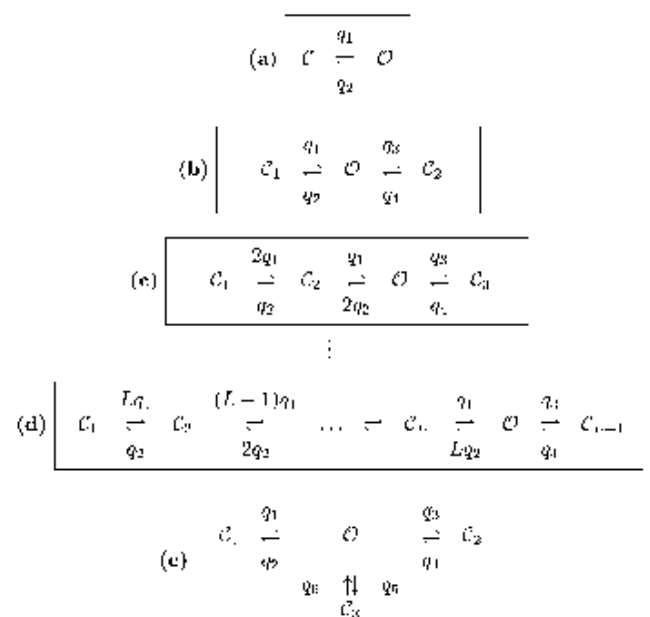
Soms is duidelijk waar dit aan ligt, zoals wanneer we in figuur 6d in het model met L = 4 een extra gesloten toestand aan de linkerkant toevoegen. In dit geval lijkt het model iets beter te worden, maar de verbetering is marginaal want de geschatte verblijftijd in die extra gesloten toestand blijkt heel klein te zijn. De extra toestand is daarom misschien niet werkelijk aanwezig. In andere gevallen is niet duidelijk waarom de geschatte a-posteriorikansen op elkaar lijken. Het zou kunnen liggen aan misspecificatie van sommige delen van het model of van de a-prioriverdeling, aan te geringe omvang van de dataset, aan gebruik van de verkeerde samplefrequentie of aan het niet lang genoeg laten lopen van de MCMC-procedure. Bovendien rijst de vraag of het theoretisch wel mogelijk is, dat een Bayesiaanse a-posterioriverdeling voor hidden Markovmodellen eenduidig de beste onderliggende Markovke-

ten voor het gating-mechanisme kan aanwijzen. Op dit gebied zijn dan ook nog veel wiskundig interessante vragen onbeantwoord.

Toch zijn we met onze analyse van de K⁺-outward-rectifier data wel wijzer geworden. We rapporteren hierover in De Gunst en Schouten (2002). Ten eerste zijn sommige van de kandidaat-modellen op grond van onze analyse niet geschikt gebleken. Ten tweede hebben we voor de wel geschikte modellen inzicht gekregen in de mogelijke range van de parameterwaarden. Het aardige is, dat een van de modellen die het best uit de analyse komen, het model is met een open toestand met aan de ene kant vier identieke langzame gesloten toestanden en aan de andere kant een snelle gesloten toestand, figuur 6d met L = 4 en met q₃ en q₄ vele malen groter dan q₁ en q₂.

De schatting van vier langzame identieke, gesloten toestanden komt overeen met de resultaten van een studie van metingen aan hele cellen van hetzelfde type kanaal (S.A. Vogelzang, persoonlijke communicatie). Deze studie is gebaseerd op een Hodgkin-Huxley-analyse (zie Van Duijn (1993)). Met onze methode vinden we echter ook de snelle gesloten toestand, waarin het kanaal telkens slechts zeer kort verblijft en die verantwoordelijk zal moeten zijn voor het typische flikkergedrag van dit kanaal. De biologische vraag is nu wat ten grondslag ligt aan het verschil tussen de twee typen gesloten toestanden. Ten derde hebben we voor de meest geschikte modellen inzicht gekregen in wat op basis van het model de meest waarschijnlijke toestand van het kanaal was op elk tijdstip van het experiment. Daarmee kunnen we een reconstructie maken van het meest waarschijnlijke open-geslotenproces. In figuren 5b en 5d zien we het op grond van het hidden Markovmodel met onderliggend Markovketenmodel 6d (L = 4) geschatte open-geslotenproces behorend bij de opname in figuur 5a, respectievelijk, 5c.

Een volgende stap is om voor de verschillende parameters de afhankelijkheid van de spanning te onderzoeken. Een eerste aanzet hiertoe is ook te vinden in De Gunst en Schouten (2002). Daarnaast kunnen we opnamen met meerdere kanalen of metingen aan hele cellen gaan



Figuur 6 Markov ketens die kandidaatmodellen zijn voor het gating mechanisme; C staat voor een gesloten en O voor een open toestand, de q's symboliseren de overgangsnelheden.

analyseren, waardoor we inzicht zouden kunnen krijgen in mogelijke afhankelijkheden tussen het kinetisch gedrag van de verschillende kanalen. Hodgkin-Huxley-analyse is gebaseerd op onafhankelijkheid van de verschillende kanalen. Dit is waarschijnlijk geen terechte aanneme. Voorts zou het interessant zijn onze methoden toe te passen op opnamen afkomstig van andere kanalen.

Conclusie

Twee heel verschillende onderzoeksprojecten zijn beschreven waarin wiskunde biologische kennis een stapje verder heeft gebracht en biologische vragen tot ontwikkeling van nieuwe wiskunde hebben geleid. Ook in de komende decennia zal er veel interessants te doen zijn op het grensvlak tussen wiskunde en biologie, met name op het terrein van de

statistiek en de moleculaire biologie. Toch wordt er op dit moment in Nederland nog niet heel veel samengewerkt door wiskundigen en biologen, hoewel de combinatie dus zeker niet nieuw is. Laten we daarom hopen dat het niet blijft bij het propageren van multidisciplinair onderzoek alleen, maar dat biologen en wiskundigen elkaar daadwerkelijk zullen opzoeken om samen te werken.

Verantwoording

Het hier beschreven onderzoek is verricht in samenwerking met Georg Luebeck, Suresh Moolgavkar, Anup Dewanji en Bettina Grasl-Kraupp (carcinogenese) en met Barry Schouten, Sake Vogelzang en Hans Künsch (ionkanalenkinetiek). De auteur dankt Rob Bogers en André Ran voor kritische lezing van een eerdere versie van dit artikel.

Referenties

- Baum, L.E. en Petrie, T. (1966), Statistical inference for probabilistic functions of finite state Markov chains, *Ann. Math. Statist.* 37: 1554–1563.
- Bernardo, J.M. en Smith, A.F.M. (1994), *Bayesian theory*, Wiley, Chichester.
- Brockwell, P.J. en Davis, R.A. (1996), *Introduction to time series and forecasting*, Springer, New York.
- Chung, S.H., Moore, J.B., Xia, L., Premkumar, L.S. en Gage, P.W. (1990), Characterization of single channel currents using digital signal processing techniques based on hidden Markov models, *Phil. Trans. R. Soc. Lond. B* 329:265–285.
- Colquhoun, D. en Hawkes, A.G. (1995). The principles of stochastic interpretation of ion-channel mechanisms, in: B. Sackman en E. Neher (eds.) 1995, *Single-channel recording*, 2nd ed. Plenum Press, New York.
- De Gunst, M.C.M. (1989), A random model for plant cell population growth. *CWI Tract* 58, Centre for Mathematics and Computer Science, Amsterdam.
- De Gunst, M.C.M., Dewanji, A. en Luebeck, E.G. (2003), Exploring heterogeneity in tumor data using Markov chain Monte Carlo, *Statistics in Medicine*.
- De Gunst, M.C.M., Künsch, H.-R. en Schouten, J.G. (2001), Statistical analysis of ion channel data using hidden Markov models with correlated state-dependent noise and filtering, *J. Amer. Statist. Assoc.* 96: 805–815.
- De Gunst, M.C.M. en Luebeck, E.G. (1994), Quantitative analysis of two-dimensional observations in the presence or absence of malignant tumors, *Math. Biosci.* 119: 5–34.
- De Gunst, M.C.M. en Luebeck, E.G. (1998), A method for parametric estimation of the number and size distribution of cell clusters from observations in a section plane, *Biometrics* 54: 100–112.
- De Gunst, M.C.M. en Schouten, J.G. (2002), Parameter estimation and model selection for recordings of the outward-rectifier in barley leaf, *Aangeboden voor publicatie*.
- De Gunst, M.C.M. en Schouten, J.G. (2003), Model selection for hidden Markov models of ion channel data using reversible jump MCMC, *Bernoulli*.
- Feller, W. (1991), *An introduction to probability theory and its applications*, Vol. I, 3rd ed., South Asia Books.
- Gelfand, A.E. en Smith, A.F. (1990), Sampling-based approaches to calculating marginal densities, *J. Amer. Statist. Assoc.* 85: 398–409.
- Gilks, W.R., Spiegelhalter, D.J. en Richardson, S. (1996), *Markov chain Monte Carlo in practice*, Chapman and Hall, London.
- Grasl-Kraupp, B., Luebeck, E.G., Wagner, A., Löw-Baselli, A., De Gunst, M.C.M., Waldhör, T., Moolgavkar, S.H. en Schulte-Hermann, R. (2000), Quantitative analysis of tumor initiation in rat liver: role of cell replication and cell death (apoptosis), *Carcinogenesis* 21: 1411–1421.
- Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika* 82: 711–732.
- Gundersen, H.J.G., Bagger, P., Bendtsen, T.F., Evans, S.M., Korbo, L., Marcussen, N., Møller, A., Nielsen, K., Nyengaard, J.R., Pakkenberg, B., Sørensen, F.B., Vesterby, A. en West, M.J. (1988), The new stereological tools: Disector, fractionator, nucleator, and point sampled intercepts and their use in pathological research and diagnosis, *Acta Pathologica Microbiologica et Immunologica Scandinavica*, 96: 857–881.
- Hastings, W.K. (1970), Monte Carlo sampling using Markov chains and their applications, *Biometrika* 57: 97–109.
- Hille, B. (1992), *Ionic channels of excitable membranes*, 2nd ed. Sunauer, Sunderland Mass.
- Karlin, S. en Taylor, H.M. (1997), *A first course in stochastic processes*, 2nd ed. Academic Press, New York.
- Kass, R.E. en Raftery, A.E. (1995), Bayes factors and model uncertainty, *J. Amer. Statist. Assoc.* 90: 773–795.
- Knudson, A.G. (1971), Mutation and cancer: statistical study of retinoblastoma, *Proc. Natl. Acad. Sci. USA* 68: 820–823.
- Kopp-Schneider, A. (1997), Carcinogenesis models for risk assessment, *Statist. Methods Med. Res.* 6: 317–340.
- Luebeck, E.G. en De Gunst, M.C.M. (2001), A stereological method for the analysis of cellular lesions in tissue sections using 3-dimensional cellular automata, *Math. and Comp. Modelling* 33: 1387–1400.
- Luebeck, E.G., Moolgavkar, S.H., Buchmann, A. en Schwarz, M. (1991), Effects of polychlorinated biphenyls in rat liver: quantitative analysis of enzyme altered foci, *Toxicol. Appl. Pharmacol.* 111: 469–484.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. en Teller, E. (1953), Equation of state calculations by fast computing machine, *J. Chem. Phys.* 21: 1087–1091.
- Moolgavkar, S.H., Luebeck, E.G., De Gunst, M.C.M., Port, R.E. en Schwarz, M. (1990), Quantitative Analysis of enzyme-altered foci in rat hepatocarcinogenesis experiments, I, Single agent regimen, *Carcinogenesis* 11: 1271–1278.
- Moolgavkar, S.H. en Venzon, D.J. (1979), Two-event models for carcinogenesis: incidence curves for childhood and adult tumors, *Math. Biosci.* 47: 55–77.
- Rabiner, L.R. en Juang, B.H. (1986), An introduction to hidden Markov models, *IEEE ASSP Magazine* 3: 4–16.
- Rice, J.A. (1995), *Mathematical Statistics and Data Analysis*, 2nd ed., Duxbury Press, Belmont, California.
- Sackman B. en Neher, E. (eds.) (1995), *Single-channel recording*, 2nd ed., Plenum Press, New York.
- Schouten, J.G. (2000), *Stochastic modelling of ion channel kinetics*, proefschrift, Vrije Universiteit Amsterdam.
- Stoyan, D., Kendall, W.S. en Mecke, J. (1995), *Stochastic Geometry and Its Applications*, 2nd ed. Wiley, New York.
- Toffoli, T. en Margolus, N. (1989), *Cellular automata machines: a new environment for modeling*, MIT Press, Boston.
- Van Duijn, B. (1993), Hodgkin-Huxley analysis of whole-cell outward rectifying K^+ currents in protoplasts from tobacco cell suspension cultures, *J. Membrane Biology* 132: 77–85.
- Van Leeuwen, I.M.M. and Zonneveld, C. (2001), From exposure to effect: a comparison of modeling approaches to chemical carcinogenesis, *Mutation Research* 489: 17–45.
- Wicksell, S.D. (1925), The corpuscle problem. A mathematical study of a biometric problem, *Biometrika* 7: 84–99.