Statistical methods for analysing complex genetic traits

El Galta, Rachid

CHAPTER 9

Samevatting

Veel aandoeningen zijn het resultaat van meerdere genetische en/of omgevingsfactoren. In de genetische epidemiologie noemt men deze aandoeningen gecompliceerde erfelijke aandoeningen. Genen die leiden tot het verkrijgen van dergelijke gecompliceerde erfelijke aandoening zijn vaak moeilijk op te sporen. Zij hebben vaak lage *penetranties*. Dat wil zeggen dat sommige patiënten de genvariant wel dragen maar de ziekte niet ontwikkeld hebben. Patiënten met dezelfde ziekte kunnen ook verschillende combinaties van ziekteveroorzakende genen dragen. Men noemt dit fenomeen een heterogeniteit effect. Samenvattend is het onderliggende genetische mechanisme van een gecompliceerde erfelijke aandoening meestal onbekend, en daarom is het moeilijk te ontdekken. In tegenstelling tot gecompliceerde erfelijke ziektes, zijn de erfelijke eigenschappen van veel Mendeliaanse ziektes gemakkelijk in de kaart te brengen. Mendeliaanse ziektes worden veroorzaakt door één enkel gen en hebben daarom een duidelijk overervingpatroon. Methoden om de onderliggende genetische achtergrond van zulke ziektes te bestuderen zijn wel in de literatuur beschreven. Echter, het opsporen van gecompliceerde ziekteveroorzakende genen vergt een andere benadering en vraagt om nieuwe statistische methoden. Een veel gebruikte strategie om gecompliceerde erfelijke aandoeningen in kaart te brengen is te beginnen met een scan van het gehele genoom waarbij genetische familie verbanden geanalyseerd worden (linkage analyses.) Dit kan helpen bij het identificeren van chromosoom-regio's die mogelijk ziekteveroorzakende genen bevatten. Vervolgens voert men een associatie studie uit om dergelijke kandidaat regio's te verfijnen. Binnen deze methodologie stelt dit proefschrift zich tot doel om nieuwe statistische methoden te ontwikkelen en te evalueren, die geschikt kunnen zijn voor het analyseren van genetische gegevens van patiënten met een gecompliceerde erfelijke aandoening in zowel families als populaties.

Voordat men een genetische studie kan opzetten, moet men eerst nagaan of er wel genetische factoren betrokken zijn bij de etiologie van de bestudeerde aandoening. Dit kan gedaan worden met behulp van een analyse

van de familiale aggregatie van de aandoening. Hiervoor kan men aselecte steekproeven van families verzamelen. Echter, meestal worden families van patiënten met de aandoening (probands) opzettelijk geselecteerd. Dit laatste wordt vooral gedaan om de statistische power te verhogen.

In **hoofdstuk 2** hebben wij een nieuwe statistische methode ontwikkeld om de aanwezigheid van clustering van een aandoening binnen families van probands te toetsen. Deze methode houdt rekening met de selectieprocedure. De familiale correlatie wordt gemodelleerd aan de hand van een stochastisch effect met verwachting nul, onbekende variantie en een correlatie die van tevoren aangegeven moet worden. Toetsen op de afwezigheid van een correlatie structuur is equivalent aan toetsen of de variantie van het stochastische effect gelijk aan nul is. De methode doet geen specifieke aanname over de verdeling van het stochastisch effect. Om familiale clustering te toetsen kan men de correlatie bepalen als een functie van de graad van de verwantschap tussen familieleden (kinship coefficient). De methode kan ook gebruikt worden om de genetische correlatie te toetsen die gedeeltelijk te wijten aan een bepaalde locus. Verder hebben we de methode geïllustreerd aan de hand van genetische gegevens van families van probands met ouderdomssuikerziekte.

In **hoofdstuk 3** hebben wij een niet-parametrische methode ontwikkeld om de genetische linkage globaal te toetsen over het hele genoom of op een gedeelte van een chromosoom. Deze methode is gebaseerd op een vergelijking van het aantal IBD allelen van een marker dat twee familie leden met de aandoening delen met het aantal dat ze zouden delen per toeval. IBD is de Engelse afkorting voor *identical by descent*. Het aantal IBD allelen dient als een maat van genetische gelijkenis tussen twee personen. De methode toetst alle markers tegelijkertijd zodat het meervoudige toetsingsprobleem vermeden wordt. Dit wordt bereikt door het optellen van de voorwaardelijke likelihood functies gegeven dat een marker de ziekteveroorzakende gen bevat. Wij hebben twee toets statistieken afgeleid, namelijk de likelihood ratio toets en de score toets. Verder hebben wij de werking van deze twee toetsen bestudeerd aan de hand van een simulatiestudie. Twee modellen werden beschouwd: (1) *één-locus* model en (2) *twee-locus* model. Het blijkt dat de likelihood ratio toets het beste presteert wat betreft van onderscheidend vermogen (power), terwijl de scoretoets alleen geschikt is voor het vinden van kandidaat-regio's of voor situaties waarin het effect van het ziekteveroorzakende gen heel klein is.

Hoofdstukken 4, 5 and 6 beschrijven statistische methoden voor genetische associatie analyses in de populatie. Hierbij vergelijkt men de allelfrequenties tussen een groep patiënten en een willekeurige groep uit de populatie. De

nadruk in deze hoofdstukken ligt op markers met meerdere allelen of haplotypen uit meerdere *enkelvoudige nucleotide polymorphismes* (SNPs). De methodes gebruiken een *semi-Bayesiaanse* aanpak waarbij de totale likelihood de som is van de gewogen voorwaardelijke likelihood functies gegeven dat een allel (haplotype) geassocieerd is met de desbetreffende ziekte. Voor deze aanpak heeft Terwilliger (1995) de gewichten gelijk gesteld aan de allelfrequenties en de corresponderende likelihood ratio (TLR) gebruikt als toets statistiek.

In **hoofdstuk 4** hebben wij dezelfde likelihood as Terwilliger (1995) gebruikt en daaruit een scoretoets afgeleid. De scoretoets is eenvoudig en, in tegenstelling tot de likelihood ratio toets, doet hij geen aanname over het aantal geassocieerde allelen. De nulverdeling van de toets kan gevonden worden met behulp van Monte Carlo simulaties. Verder hebben wij deze toets analytisch vergeleken met de bekende Chi-kwadraat toets voor cruistabellen. Het blijkt dat de scoretoets het beter doet dan Chi-kwadraat toets wanneer het geassocieerde allel frequent voorkomt. Aan de hand van een simulatie hebben wij ook de power van de scoretoets vergeleken met andere bestaande toetsen, inclusief de Chi-kwadraat toets. Wij hebben zowel markers met één geassocieerd allel als met twee geassocieerde allelen beschouwd. Het blijkt dat de scoretoets gemiddeld de beste power heeft. Ter illustratie hebben wij de scoretoets toegepast op echte data.

Hoofdstuk 5 beschrijft de resultaten van toepassingen van de scoretoets uit hoofdstuk 4, TLR and Chi-kwadraat op kandidaat- regio's. De kandidaatregio's zijn geïdentificeerd met behulp van linkage analyses op een aantal chromosomen. De data waren afkomstig uit het *Genetic Association Workshop* 14 gesimuleerde microsatellieten probleem (Bailey-Wilson et al., 2005; Greenberg et al., 2005). De bestudeerde ziekte was "Kofendrerd Personality Disorder" (KPD), een fictieve aandoening. Alle toegepaste toetsen duiden aan dat er mogelijk een associatie bestaat tussen KPD en de microsatelliet marker D03S0127. Verdere associatie-analyses tussen KBD en 20 SNPs rond marker D03S0127 laten zien dat er een sterke associatie is tussen KBD en de SNP B03T3057 wat eigelijk ligt naast de echte ziekte locus in het end van chromosoom 3. (Greenberg et al., 2005). De analyses werden uitgevoerd zonder enige voorkennis van de antwoorden.

Hoofdstuk 6 presenteert een generalisatie van de methode beschreven in hoofdstuk 4. Wij hebben de gewichten als allel- (haplotype-) frequenties in de totale likelihood vervangen door constante gewichten. In deze nieuwe aanpak hoeven de gewichten dus niet geschat te worden uit de data. Voor deze likelihood hebben wij een nieuwe scoretoets afgeleid. De scoretoets is eenvoudig en

heeft een normale verdeling. Het specificeren van de gewichten kan gedaan worden op basis van de voorkennis dat onderzoekers bijvoorbeeld hebben opgedaan uit eerder studies. Echter, men kan ook niet-informatieve (gelijke) gewichten gebruiken wanneer er geen op voorhand informatie beschikbaar is. Dat wil zeggen dat alle allelen (haplotypes) a-priori een even grote kans hebben om geassocieerd te zijn met de ziekte. Voor gelijke gewichten hebben wij verder een uitgebreide simulatie studie uitgevoerd om de werking van deze nieuwe toets in vergelijking met Chi-kwadraat en TLR toets te bestuderen zowel onder het nulmodel als onder het alternatieve model. Wij hebben alleen markers (haplotypes) met één geassocieerde variant beschouwd. In het algemeen presteert de scoretoets het beste. De werking van scoretoets is met succes geïllustreerd in de context van haplotype analyses aan de hand van data van drie kandidaatgenen uit *Leiden Thrombophilia Studie* . De analyses zijn uitgevoerd onder de veronderstelling van perfecte informatie omtrent *haploype phase* .