Analysis of metabolomics data from twin families

Draisma, H.H.M.

Samenvatting

Dit proefschrift beschrijft de resultaten van verschillende analyses die uitgevoerd kunnen worden in het kader van tweelingenstudies op basis van *metabolomics* data. De tweelingenstudie is een gevestigde methode om te schatten of verschillen tussen personen in meetbare eigenschappen hoofdzakelijk toe te schrijven zijn aan genetische invloeden, danwel aan verschillen in omgeving. *Metabolomics* is een betrekkelijk jonge tak binnen de "*omics*" wetenschappen, die tot doel heeft een uitputtend overzicht te geven van de stoffen (metabolieten) die betrokken zijn bij biochemische processen in biologische systemen. Een onderdeel van *metabolomics* is het meten van de concentraties of onderlinge verhoudingen in concentraties van deze metabolieten in lichaamsvloeistoffen zoals bloed en urine. Centraal in dit proefschrift staan de ontwikkeling en de toepassing van methodes voor analyse van de data die voortkomen uit dergelijke metingen in het kader van (tweelingen)familiestudies. Zodoende draagt dit proefschrift bij aan het ophelderen van de bijdrage van genetische en omgevingsinvloeden aan individuele verschillen in metabolietconcentraties in lichaamsvloeistoffen.

In Hoofdstuk 1 van dit proefschrift wordt een algemene inleiding gegeven in *metabolomics* en tweelingen- en familiestudies. Uiteengezet wordt welke rol familiestudies, en in het bijzonder studies van tweelingen en hun naaste familieleden, hebben voor het bestuderen van de factoren die ten grondslag liggen aan de individuele verschillen voor meetbare eigenschappen die in hun waarde geleidelijk variëren tussen personen. Vanwege hun belangrijke rol in dit proefschrift worden twee methodes geïntroduceerd die kunnen worden gebruikt voor statistische analyse binnen dergelijke studies. De eerste van deze technieken, *structural equation modeling* (SEM), gaat uit van een model dat gebaseerd is op een hypothese met betrekking tot de oorzaken van variatie binnen en tussen verschillende meetbare eigenschappen. In een dergelijk model zijn de bijdragen van de verschillende oorzaken van meetbare variatie opgenomen als parameters

die vrij in waarde kunnen variëren. De parameterwaarden die het beste bij de meetgegevens aansluiten, kunnen worden aangenomen als schattingen voor de waarden van de betreffende parameters in de onderzochte populatiesteekproef.

De tweede techniek voor het bestuderen van de onderlinge verschillen in meetbare eigenschappen die besproken wordt in Hoofdstuk 1 is hiërarchische clusteranalyse. Met behulp van deze techniek kan een overzicht worden verkregen van de onderlinge overeenkomsten tussen variabelen (bijvoorbeeld, metabolieten) of tussen objecten (bijvoorbeeld, proefpersonen) op basis van meetgegevens voor meerdere eigenschappen gemeten in een steekproef. In Hoofdstuk 1 wordt uiteengezet dat in dit proefschrift deze techniek op twee verschillende manieren gebruikt wordt om inzicht te geven in de genetische factoren die ten grondslag liggen aan onderlinge verschillen in meetbare eigenschappen. Voorts wordt in Hoofdstuk 1 een perspectief geschetst hoe studies zoals beschreven in dit proefschrift een overzicht kunnen geven van de genetische en omgevingsinvloeden op de concentraties van verschillende elementen van biologische systemen (bijvoorbeeld, gentranscripten, enzymen en metabolieten) afzonderlijk en in hun samenhang. Tot slot wordt in Hoofdstuk 1 betoogd dat studies op basis van meetgegevens verkregen in bijvoorbeeld tweelingenfamilies, de interpretatie van genoom-brede associatiestudies kunnen verbeteren en onder andere daarmee een bijdrage kunnen leveren aan de opheldering van zogenaamde complexe aandoeningen.

Hoofdstuk 2 beschrijft de onderlinge verschillen en overeenkomsten in lipidenprofielen zoals gemeten in bloedplasma tussen (voornamelijk ééneiige) tweelingen en hun niet-tweelingbroers en -zussen. Deze lipidenprofielen werden verkregen met één van de meest gebruikte meetmethodes binnen *metabolomics*, te weten vloeistofchromatografie gekoppeld aan massaspectrometrie (LC–MS). Bij deze techniek worden de componenten in het onderzochte monster eerst gescheiden op een chromatografische kolom op basis van hun verschillen in fysisch-chemische eigenschappen, en vervolgens gedetecteerd met een massaspectrometer. In de studie zoals beschreven in Hoofdstuk 2 werden in het bloedplasmamonster van iedere proefpersoon met LC–MS relatieve concentraties gemeten van in totaal 61 verschillende lipiden uit vijf verschillende lipidenklassen. De gemeten lipiden zijn betrokken bij een breed scala van fysiologische en pathofysiologische processen, waaronder signaaltransductie, ontstekingsreacties en energiehuishouding.

Hiërarchische clusteranalyse werd in Hoofdstuk 2 gebruikt om de proefpersonen te groeperen op basis van hun onderlinge verschillen en overeenkomsten in plasmalipidenprofiel. Het resultaat van deze groepering suggereerde dat mannelijke en vrouwelijke proefpersonen verschillende lipidenprofielen hadden. Verdere analyses van de resultaten van de hiërarchische clusteranalyse onderbouwden de hypothese dat in het algemeen, overeenkomsten in genetische en omgevingsinvloeden bijdragen aan overeenkomsten in lipidenprofielen tussen personen. Echter, in het onderzoek zoals beschreven in Hoofdstuk 2 werden ook aanwijzingen gevonden dat bepaalde omstandigheden, waaronder recente ziekte, samengaan met veranderingen in het lipidenprofiel zoals gemeten in

bloedplasma.

Het onderscheidend vermogen van statistische toetsen wordt vergroot door het aantal waarnemingen te vergroten op basis waarvan getoetst wordt. In dit verband beschrijft Hoofdstuk 3, een techniek genaamd "*quantile equating*", die het mogelijk maakt meetgegevens te combineren die verkregen zijn met dezelfde semi-kwantitatieve analytisch-chemische methode, maar bijvoorbeeld op verschillende tijdstippen binnen een grote studie. In dit hoofdstuk wordt beargumenteerd dat het gebruik van dergelijke datavoorbewerkingstechnieken noodzakelijk kan zijn vanwege praktisch onvermijdbare kleine verschillen in analytische methodologie tussen 'blokken' van metingen. De succesvolle toepassing van de in dit hoofdstuk gentroduceerde methode wordt gedemonstreerd aan de hand van meetgegevens verkregen met LC–MS metingen van relatieve concentraties van lipiden in bloedplasma, en met metingen van protonenkernspinresonantie (^{1}H NMR) in bloedplasma en in urinemonsters. Alle waterstofhoudende moleculen in een monster dragen bij aan het met ^{1}H NMR gedetecteerde signaal, en daarmee bevatten ^{1}H NMR data informatie over de concentraties van metabolieten behorend tot verschillende klassen.

De met behulp van de in Hoofdstuk 3 beschreven methode gecombineerde LC–MS data, werden vervolgens gebruikt voor de studie zoals beschreven in Hoofdstuk 4. Deze samengestelde dataset bevatte meetgegevens voor ééneiige tweelingen en hun niet tweelingbroers en -zussen zoals beschreven in Hoofdstuk 2, en gegevens voor twee-eiige tweelingen en hun niet tweelingbroers en -zussen. Evenals in Hoofdstuk 2 werd hiërarchische clusteranalyse toegepast om de onderlinge variatie in lipidenprofielen tussen personen in kaart te brengen. De aanwezigheid van twee-eiige tweelingen in deze studie, maakte het mogelijk om de invloed van gedeelde omgevingsinvloeden op overeenkomsten in deze profielen beter vast te kunnen stellen. Tevens werden, evenals in Hoofdstuk 2, aanwijzingen gevonden dat bijvoorbeeld ziekte in het recente verleden samen zou kunnen gaan met veranderingen in het lipidenprofiel in bloedplasma. Daarnaast bevestigden de resultaten van de hiërarchische clusteranalyse dat toepassing van een methode zoals beschreven in Hoofdstuk 3 noodzakelijk was geweest om de meetgegevens verkregen in verschillende blokken samen te kunnen voegen.

Deze zelfde gecombineerde meetgegevens op basis van LC–MS analyse van lipiden in bloedplasma, evenals de samengevoegde gegevens verkregen met ^{1}H NMR metingen in bloedplasma, werden gebruikt voor de analyses zoals beschreven in Hoofdstuk 5. Allereerst werden in dit onderzoek met SEM de relatieve bijdragen geschat van genetische invloeden en van omgevingsinvloeden aan de verschillen tussen personen in de concentraties gemeten voor elke afzonderlijke variabele. Dergelijke analyses van de triglyceriden gemeten met LC–MS lieten een consistent patroon zien, waarbij zowel het aantal koolstofatomen als het aantal dubbele bindingen in de vetzuurstaarten van belang leek voor de erfelijkheid van de gemeten concentraties.

Vervolgens werden in een bivariate SEM-analyse, voor elke paarsgewijze combinatie van variabelen gemeten met elk van beide analytische methoden,

de genetische correlaties tussen variabelen berekend. Hiërarchische clusteranalyse werd gebruikt om de patronen in deze genetische correlaties voor zowel de LC–MS gegevens als de ^{1}H NMR gegevens te onderzoeken. Hierbij viel op dat de positieve correlatie tussen verschillende lipiden in bloedplasma behorend tot dezelfde lipidenklasse, samen leek te hangen met gedeelde erfelijke invloeden tussen deze lipiden. Binnen de variabelen gemeten met ^{1}H NMR werden grote verschillen in de genetische correlaties waargenomen, die te maken zouden kunnen hebben met het feit dat met de gebruikte ^{1}H NMR methode metabolieten van verschillende metabolietklassen waargenomen kunnen worden.

Enkele conclusies en aanbevelingen voor verder onderzoek naar aanleiding van dit proefschrift worden beschreven in Hoofdstuk 6. Één van de conclusies is dat methodes toegepast binnen *microarray* onderzoek voor het combineren van data gemeten voor dezelfde gentranscripten maar in verschillende personen, mogelijk toepasbaar zijn binnen *metabolomics* voor het combineren van data gemeten voor dezelfde metabolieten maar in verschillende personen. Deze methodes kunnen daarmee een aanvulling vormen op de methode zoals beschreven in Hoofdstuk 3 van dit proefschrift. Ook wordt in Hoofdstuk 6 betoogd dat toepassing van de relatief 'hypothese-vrije' methode voor clustering van proefpersonen zoals beschreven in Hoofdstukken 2 en 4 resultaten gaf die consistent waren met een vooraf opgestelde hypothese. Anderzijds werd de 'hypothese-gedreven' methode SEM in Hoofdstuk 5 op een relatief hypothese-vrije manier toegepast, wat evenwel eveneens resultaten gaf die consistent waren met de biologische achtergrond van de data. Een mogelijke toepassing van de resultaten van hiërarchische clusteranalyse van proefpersonen voor het vinden van voor SEM relevante covariaten wordt beschreven.

Tot slot wordt aangegeven dat methodes om ziekte op te sporen door de analyse van afstanden tussen personen, zoals bijvoorbeeld beschreven in Hoofdstukken 2 en 4 van dit proefschrift, verder onderzoek verdienen vanwege het mogelijk hogere onderscheidend vermogen ten opzichte van methodes gebaseerd op analyse van afzonderlijke variabelen.