Statistical methods for the analysis of complex omics data

Tissier, R.

Samenvatting

Deze dissertatie richt zich op de ontwikkeling van nieuwe statistische methodes waarbij rekening wordt gehouden met afhankelijkheidsstructuren in omics-datasets en met de modellering van deze structuren. Statistische modellering van deze structuren kan leiden tot verdere verbetering van onze kennis van biologische mechanismen. Door rekening te houden met de structuur zijn wij mogelijk beter in staat om ziekten te voorspellen. In hoofdstuk 1 wordt de meest gebruikelijke mate van afhankelijkheid beschreven, namelijk de correlatiecoëfficiënt van Pearson. Verder wordt een algemene inleiding gegeven over verschillende afhankelijkheidsstructuren waarmee iemand bij het bestuderen van omics-datasets mogelijk te maken krijgt: afhankelijkheden tussen personen, gemeten resultaten en omics-eigenschappen. Voor elk van deze afhankelijkheidsniveaus worden de uitdagingen beschreven die iemand kan tegenkomen en de daarvoor meestal gebruikte methodes.

In hoofdstuk 2 en 3 worden methodes beschreven voor de analyse van secundaire fenotypen in onderzoek naar geverifieerde families. In hoofdstuk 2 wordt ingegaan op een nieuwe aanpak voor het analyseren van het secundaire fenotype voor een opzet voor familieonderzoek met meerdere casussen, waarvoor families worden geselecteerd waarin minimaal een specifiek aantal casussen voorkomt. De voorgestelde methode wordt onderbouwd aan de hand van een voorbeeld op basis van gegevens van het onderzoek Leiden Lang Leven, een familieonderzoek aan de hand van meerdere casussen, waarin de veroudering bij mensen (primair fenotype) wordt onderzocht. Hier werd een inschatting gemaakt van de parameters die zorgen voor de verbanden tussen triglyceridespiegels en glucose (secundaire fenotypen) en genetische markers. In hoofdstuk 3 worden methodes beschreven die binnen de literatuur zijn terug te vinden ten aanzien van secundairefenotypeanalyse voor een onderzoeksopzet op basis van de familie van proefpersonen. Bij deze opzet worden familieleden van specifieke proefpersonen (vaak casussen met de primaire resultaten) meegenomen in het onderzoek. De prestaties van de beschikbare methodes worden vergeleken met onze methode, die beschreven staat in hoofdstuk 2. De analyse van werkelijke gegevens in dit hoofdstuk maakt onderdeel uit van het familieonderzoek naar socialeangststoornissen (Social Anxiety Disorder, SAD) en richt zich op het vaststellen van kandidaat- endofenotypen van SAD.

Hoofdstuk 2 omschrijft een benadering voor het verkrijgen van onvertekende associatieve schattingen tussen secundaire fenotypen en biomarkers en onvertekende erfelijkheidsschattingen voor secundaire fenotypen. Deze methode biedt ruimte voor het vaststellingsproces en zorgt voor expliciete modellering van de familierelaties. Om dit te be-

reiken maken wij bij onze benadering gebruik van retrospectieve waarschijnlijkheid, ten behoeve van modellen met gemengde effecten. De benadering op basis van retrospectieve waarschijnlijkheid corrigeert automatisch voor de vaststelling. De willekeurige effecten zorgen voor modellering van de familierelaties. Om de associatie tussen de primaire fenotypen (binaire variabele) van het gemengde type en de secundaire fenotypen (continue variabele) te kunnen bepalen wordt gebruikgemaakt van een multivariaat probitmodel. Door maximalisatie van de retrospectieve log-waarschijnlijkheid kunnen er schattingen worden gedaan.

Een belangrijke empirische bevinding is dat de erfelijkheidsschattingen voor de secundaire trekken sterk kunnen worden onderschat, tenzij rekening wordt gehouden met de wijze van monstername. Uit uitgebreide simulaties is gebleken dat de hier gepresenteerde methode de fout van type 1 op een nominaal niveau houdt en zorgt voor nauwkeurige schattingen, ongeacht de prevalentie van de ziekte, de sterkte van de associatie tussen de genetische varianten en het primaire fenotype, en ongeacht het vaststellingsmechanisme. Momenteel is een belangrijke beperking van onze benadering de aanwezigheid van multivariate integralen, waarvan de berekening veel tijd kost, vooral als er sprake is van een grote stamboom.

In hoofdstuk 3 wordt onderzoek gedaan naar de prestaties van onze methode voor de analyse van gegevens van proefpersonenonderzoeken met familiebenadering, zoals die in hoofdstuk 2 beschreven worden. De prestaties worden vergeleken met methodes die vandaag de dag binnen de literatuur gangbaar zijn, namelijk methodes die het vaststellingsproces negeren of die gezien de secundaire fenotypen van de proefpersonen de voorwaardelijke spreiding van de waarden van de secundaire fenotypen van de families modelleren. Verder pleiten wij voor een uitbreiding van deze laatste wijze van aanpak, waarbij bij de gezamenlijke, voorwaardelijke spreiding van de primaire en secundaire fenotypewaarden van de familieleden wordt uitgegaan van de gezamenlijke spreiding van de primaire en secundaire fenotypen van de proefpersonen.

Uit uitgebreide simulaties is gebleken dat alleen de benadering op basis van retrospectieve waarschijnlijkheid die in hoofdstuk 2 werd ontwikkeld ook echt in staat is om onvertekende erfelijkheidsschattingen te krijgen van het secundaire fenotype, evenals onvertekende parameterschattingen voor de associaties tussen de secundaire fenotypen en genetische markers. Bovendien kan conditionering op de secundaire fenotypewaarden van de proefpersoon leiden tot een ernstige onderschatting van de erfelijkheid en dat kan ook de identificatie beperken van kandidaat-endofenotypen van primaire fenotypen. Alleen de benadering op basis van retrospectieve waarschijnlijkheid kon binnen de analyse van werkelijke gegevens een kandidaat-endofenotype van SAD vaststellen. Een ander belangrijk punt binnen dit hoofdstuk is dat uit alle methodes vertekende schattingen voortkomen als de informatie van de proefpersoon ontbreekt. Daarom moet de toepassing van een dergelijke onderzoeksopzet op dit moment niet worden overwogen.

In hoofdstuk 4 wordt gekeken naar het probleem van het uitvoeren van netwerkanalyse voor co-expressie van genen bij familieonderzoeken. Een grote variatie in expressieniveaus tussen families onderling zou een aanzienlijke vertekening vormen voor de

verkregen netwerkstructuur als geen rekening zou worden gehouden met de stamboomstructuur. Om dit probleem te voorkomen, stellen wij een meta-analytische benadering voor. Wij bouwen eerst het omics-netwerk op voor iedere stamboom, om zo clusters van correlerende microarray-probes vast te kunnen stellen. De eigengenen (eerste hoofdcomponent) van ieder cluster van elke stamboom worden vervolgens getest op associatie met een belangwekkend fenotype. Na bepaling van het sterkst geassocieerde cluster, worden de clusters die binnen iedere familie hiermee het meest overlappen met dit cluster gecombineerd. Ten slotte wordt het eigengen van het gecombineerde cluster getest op associatie met het fenotype. Deze methode werd gebruikt voor de analyse van de gesimuleerde dataset die beschikbaar werd gesteld voor de Genetic Analysis Workshop 18 en de prestaties ervan werden vergeleken met die van andere methodes, waaronder: enkeleprobeanalyse, waarbij de stamboomstructuur wordt genegeerd, en opbouw van het netwerk op basis van gedecorreleerde omics-variabelen. In hoofdstuk 5 en 6 presenteren wij nieuwe methodes om groeperingsinformatie in te bouwen in voorspellingsmodellen, om zo stabielere en waar mogelijk interpreteerbare modellen te kunnen krijgen. Bij alle analyses in deze hoofdstukken wordt ter illustratie gebruikgemaakt van gegevens uit het DILGOM-onderzoek (DIetary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome) en van de openbaar toegankelijke, farmacogenomische dataset van borstkankercellijnen.

In hoofdstuk 5 wordt een nieuwe aanpak gepresenteerd voor modelselectie op basis van drie stappen: opbouwen van een netwerk van omics-eigenschappen, empirische derivatie van modules met vergelijkbare eigenschappen door middel van clustering en ten slotte het opbouwen van een voorspellingsmodel, waarin de groeperingsinformatie is ingebouwd. Deze aanpak is erop gericht om problemen als gevolg van de aanwezigheid van sterke correlaties tegen te gaan. Er worden verschillende methodes afgewogen voor het uitvoeren van stap 1 en 3 van de ontwikkelde benadering. Wij vergelijken de prestaties van deze strategie door middel van simulaties met de standaard geregulariseerde regressie, zoals LASSO, Ridge-regressie en elastic net.

Uit de resultaten van de simulaties en datatoepassing blijkt dat deze aanpak leidt tot stabielere voorspellingsmodellen en dat deze wat betreft nauwkeurigheid van de voorspellingen even goed werkt als de standaard geregulariseerde regressie. Bij methodes zoals LASSO of elastic net wordt meestal een willekeurige variabele geselecteerd uit een groep sterk correlerende variabelen, wat leidt tot instabiele modellen en daardoor tot problemen met de reproductie van de resultaten. Door voorspellende prestaties van de diverse combinaties van netwerkbenaderingen en voorspellingsmodellen te vergelijken, kunnen we richtlijnen geven voor de te gebruiken combinatie van methodes. De combinatie van grafische LASSO en groeps-LASSO is de aanpak die over het algeheel gesproken de beste prestaties geeft. Bij grote datasets heeft echter WGCNA de voorkeur boven grafische LASSO, aangezien voor grafische LASSO erg intensieve berekeningen nodig zijn.

In hoofdstuk 6 onderzoeken we hoe binnen voorspellingsmodellen verschillende omicsdatasets simultaan kunnen worden gebruikt. Het combineren van verschillende omicsbronnen binnen een voorspellingsmodel is een hele uitdaging, gezien de sterke onderlinge

heterogeniteit van omics-bronnen. De gegevenssets variëren in termen van dimensionaliteit, normalisatieprocedures en foutstructuren. In dit hoofdstuk stellen wij drie strategieën voor om twee omics-bronnen binnen een voorspellingmodel te integreren. Onze specifieke voorstellen zijn: 1) stapeling van de beide omics-bronnen en toepassing van de benadering die is voorgesteld in hoofdstuk 5, 2) netwerkconstructie en clustering van beide omics-bronnen afzonderlijk en bouwen van het voorspellingsmodel, 3) netwerkconstructie en clustering van beide omics-bronnen afzonderlijk, bepaling van de correlatie tussen de clusters en tussen de omics-bronnen en inbouwen van deze informatie in het voorspellingsmodel. De voorbeelden van de gegevens in dit hoofdstuk omvatten datasets bestaande uit metabolomics en transcriptomics uit het DILGOM-onderzoek en kopienummervarianten en genexpressie van de dataset van farmacogenomische borstkankercellijnen.

De belangrijkste componenten van de door ons voorgestelde benadering zijn het bepalen van groepen van intern of onderling correlerende eigenschappen van de omics-datasets en het integreren van deze informatie door middel van een groepspenalisatiemethode. Uit simulaties blijkt dat het naïef stapelen van datasets meestal geen goede strategie is, aangezien het model meestal slechter presteert dan een model dat is gebaseerd op een enkele omics-dataset. Toevoeging van informatie over de onderlinge correlatie van de omics-datasets kan de voorspellingsnauwkeurigheid mogelijk verbeteren. Als de ruisstructuren van beide omics-bronnen verschillen, blijkt dat in termen van voorspellingsnauwkeurigheid het uitvoeren van de netwerkanalyse en clustering voor iedere omics-bron afzonderlijk robuuster is dan de stapeling van datasets.