Metagenomics : beyond the horizon of current implementations and methods

Khachatryan, L.

Samenvatting

Dankzij de ontwikkelingen in sequentietechnieken zijn metagenomen een rijke bron van informatie geworden voor vele wetenschappelijke disciplines zoals menselijke en dierlijke gezondheidszorg, ecologie, forensisch onderzoek, landbouw en voedselproductie. Een gedetailleerde analyse van metagenomische data is daarom van groot belang om alle aanwezige informatie te onthullen. Hierbij proberen wetenschappers meestal het antwoord te vinden op drie hoofdvragen:

- Welke organismen zijn aanwezig in het metagenoom?
- Wat doen ze daar?
- Wat is het verschil tussen metagenomen?

Traditioneel worden de antwoorden op de eerste twee vragen verkregen door middel van zogeheten "referentie-gebaseerde methoden", waarbij metagenomische data eerst vergeleken wordt met bekende genomen, genen of reactieketens. Een duidelijk nadeel van deze technieken is de onvolledigheid van bestaande databases: microbiële gemeenschappen bestaan veelal uit honderd tot duizenden onbekende bacteriën, omdat informatie over deze bactieriën ontbreekt is de nauwkeurigheid van referentie-afhankelijke methoden beperkt. Daarom worden referentie-vrije methoden populairder in de vergelijkende metagenomica. In mijn onderzoek tracht ik de metagenomische analyse te verbeteren in twee richtingen: mét en zonder referentie-databases (zie hoofdstuk 3 en 4).

Voor de referentie-vrije analyse van verscheidene Next Generation Sequencing datasets ontwikkelden wij een methode gebaseed op *k*-meren (kPal). We laten zien dat onze aanpak gebruikt kan worden voor twee soorten metagenomische analyse: om het niveau van verwantschap tussen twee microbiomen te kwantificeren (hoofdstuk 3), en om de genetische informatie binnen één metagenoom te classificeren (hoofdstuk 4). We hebben kPal getest op een reeks gesimuleerde metagenomen met verschillende aantallen van nauw verwante bacteriële genomen. Onze methode bleek in staat tijdelijke verandering in microbiotische compositie te detecteren. Om

te controleren of deze referentie-vrije methode het verschil tussen menselijke metagenomen kan blootleggen, hebben we onze methode ook getest op 16S metagenomen van ingewanden en de huid van verschillende testpersonen over een periode van 6 maanden. kPal kan niet alleen het verschil zien tussen de afkomst (ingewanden of huid) van het metagenoom, het kan ook het onderscheid zien tussen de verschillende testpersonen! Dit resultaat is beter dan referentie-afhankelijke methoden laten zien, die namelijk niet de huid-monsters van verschillende personen kunnen onderscheiden.
We hebben onze op *k*-meren gebaseerde methode ook toegepast om genetische sequenties te classificeren in één metagenomische dataset. Naast een aantal gesimuleerde metagenomische datasets hebben we ook data verkregen van een bioreactor microbioom met behulp van het PacBio RSII platform. We laten zien dat de *k*-mer profielen relaties kunnen onthullen tussen genetische sequenties in een enkel metagenoom, waarmee we de sequenties kunnen clusteren per soort. Deze resultaten zijn zeer belangrijk, omdat ze bewijzen dat het mogelijk is om structuren te detecteren binnen een enkel metagenoom met slechts de informatie die in het metagenoom zelf beschikbaar is. Onze referentie-vrije methode kan dus gebruikt worden voor vergelijkende metagenomica. Bovendien kunnen we sequenties in een enkel metagenoom classificeren, waardoor we de in een monster aanwezige genomen kunnen ontwaren.

Daarnaast hebben we de grenzen van referentie-afhankelijke technieken onderzocht in enkele studies (hoofdstuk 2 en 5).
Ons eerste doel was om de twee meest populaire datasoorten voor referentie-afhankelijke taxonomische profilering te vergelijken: de amplicon-gebaseerde 16S data versus de Whole Genome Sequencing (WGS; volledige genoom-sequentie) data (hoofdstuk 2). Voor dit onderzoek creërden wij een reeks kunstmatige bacteriële mengsels, elk met een andere verdeling van soorten. Deze mengsels werden gebruikt om de nauwkeurigheid van de twee datasoorten te bepalen, en om verscheidene methoden voor taxonomische classificatie te evalueren. Onze resultaten laten zien dat WGS-data veel nauwkeurigere resultaten oplevert dan 16S data. Daarmee verwerpen we dat wijdverbreide mening dat 16S data toereikend is voor de analyse van metagenomische monsters.
Tot slot hebben we de toepasbaarheid van referentie-afhankelijke methoden vergroot door een pipeline te maken die klinische monsters kan analyseren met mogelijk meer dan één pathogeen (hoofdstuk 5). Hiervoor ontwikkelden we BacTag, een gedistribueerde bioinformatica pipeline voor een snelle en accurate typering van bacteriële genen en allelen in klinische WGS-data. Het grote voordeel van onze methode bestaat uit een voorbewerkingsprocedure waarin de signatuur van elk mogelijk allel wordt geïdentificeerd en opgeslagen in een database. De daaropvolgende identificatie van allelen in een klinisch monster wordt gedaan aan de hand van deze signaturen in plaats van een traditionele uitputtende zoektocht. Omdat deze

methode ook toegepast kan worden op ongecultiveerde monsters, kan de methode goed gebruikt worden voor gevallen waar een snelle analyse van belang is.