Bedrijfstakonderzoek
BTO 2021.070 | December 2021

Monitoring van zuiveringsprocessen en distributiesystemen met metagenomics en metaproteomics

Bedrijfstakonderzoek

KWR

Bridging Science to Practice

Rapport

Monitoring van zuiveringsprocessen en distributiesystemen met metagenomics en metaproteomics

BTO 2021.070 | December 2021

Opdrachtnummer
402045/249/011

Projectmanager
Michiel Hootsmans

Opdrachtgever
BTO – Thematisch Onderzoek – Biologische Veiligheid

Auteur(s)
Michiel in 't Zandt

Kwaliteitsborger(s)
Peer Timmers i.o. Paul van der Wielen

Verzonden naar
Dit rapport is verspreid onder BTO-participanten.

Een jaar na publicatie is het openbaar.

Keywords

Metagenomics, metaproteomics, sequencing, monitoring, zuivering

Jaar van publicatie
2021

Meer informatie
Michiel in 't Zandt
T +31 (0)30 606 9502
E michiel.zandt.int@kwrwater.nl

PO Box 1072
3430 BB Nieuwegein
The Netherlands

T +31 (0)30 60 69 511
E info@kwrwater.nl
I www.kwrwater.nl

KWR

Rapport

Samenvatting

De hoofdvraag van deze bureaustudie binnen het BTO-project Moleculair Microbiologische Methoden betreft de ontwikkeling van methoden waarmee de rol van microbiologie bij zuiveringsprocessen en in het distributiesysteem kan worden onderzocht. In dit onderzoek is daarbij specifiek naar de potentie van metagenomics en metaproteomics gekeken.

Met metagenomics kan onderzoek gedaan worden naar functies van de microbiële soorten in een monster. Een grote uitdaging bij metagenoom-analyses is dat diverse analysestappen noodzakelijk zijn om van ruwe DNA-data naar nuttige informatie voor beoordeling en sturing van zuiveringsprocessen en waterdistributie te gaan. In dit onderzoek wordt gekeken naar de twee hoofdstappen in metagenoom analyse, namelijk de bioinformatica en data-interpretatie. Er is een stappenplan uitgewerkt voor de opzet van een metagenomics-workflow. Uitdagingen rondom implementatie van metagenomics liggen vooral bij de noodzaak van voldoende rekencapaciteit (een geschikte server) en de benodigde expertise in bioinformatica en data-interpretatie. Daarom is hier tevens de samenwerking met universitaire afdelingen waar deze expertise voorhanden is verkend.

Een limitatie van metagenomics is dat met deze methode alleen uitspraken kunnen worden gedaan over het metabole *potentieel*, ofwel de functies die in theorie door micro-organismen uitgevoerd kunnen worden, omdat ze de genen voor de enzymen die bij de metabole processen betrokken zijn op het genoom hebben. Voor het in kaart brengen van de metabole processen die daadwerkelijk plaatsvinden, dat wil zeggen dat de enzymen ook daadwerkelijk in de cel worden aangemaakt, kan metaproteomics gebruikt worden. Metaproteomics is een relatief nieuwe techniek die momenteel nog niet breed binnen het drinkwateronderzoek wordt toegepast en door de aard van de techniek andere uitdagingen met zich mee brengt.

Als onderdeel van deze bureaustudie naar de implementatiemogelijkheden van metaproteomics is allereerst een overzicht samengesteld voor de benodigde stappen van een metaproteoom-analyse met als doel een beter inzicht in de metaproteomics-workflow te krijgen. Hieruit blijkt dat, gezien de beperkte ervaring op drinkwatermonsters, het hiervoor het meest haalbaar is om een samenwerking op te zetten met een instituut dat zich ook op dit soort monsters richt. Hierbij kwam de Microbial Proteomics Group aan de TU in Delft naar voren. Wanneer blijkt dat metaproteomics een vaste waarde wordt in het drinkwater-onderzoek, kan op de langere termijn gekozen worden om een eigen metaproteomics-infrastructuur op te gaan zetten binnen KWR.

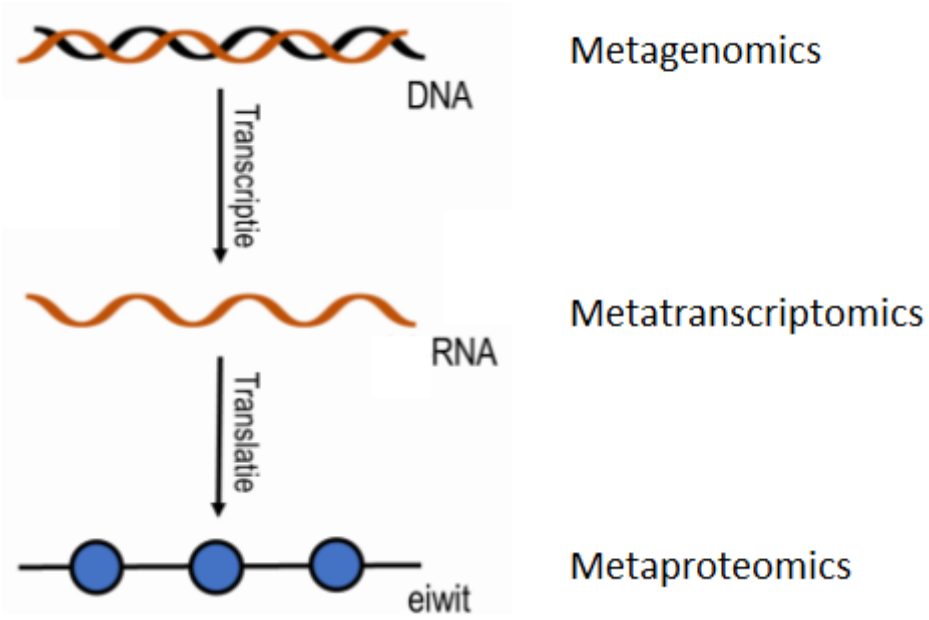
Inhoud

1 Inleiding

Met **metagenomics** is het mogelijk om informatie te krijgen over de samenstelling van de aanwezige genen in een monster, zodat kennis beschikbaar komt over de metabole potentie om biochemische processen te laten plaatsvinden (Figuur 1), zoals bijvoorbeeld de afbraakprocessen van organische of anorganische verbindingen door micro-organismen. Met **metaproteomics** wordt informatie verkregen over welke eiwitten (en dus enzymen) de microbiële populaties produceren (Figuur 1). Dit verschaft kennis over de actieve microbiële metabole processen die daadwerkelijk in het milieu plaatsvinden. Met **metatranscriptomics** wordt gekeken naar de RNA-samenstelling, wat informatie verschaft over de gen-expressie en daarmee een indicatie geeft over de potentiële activiteit van micro-organismen (Figuur 1). Gezien de lage stabiliteit van RNA en lage concentraties in drinkwater-gerelateerde monsters is deze techniek minder geschikt. Metatranscriptomics is hierdoor geen onderdeel van deze bureaustudie (zie BTO 2018.044 voor een verkenning van de toepassingsmogelijkheden).

In diverse projecten zijn eerder metagenomics-analyses verkend en uitgevoerd maar bleek generatie, verwerking en interpretatie van data met de juiste bioinformatica-tools niet optimaal waardoor de kracht van deze analyses niet volledig kon worden benut (BTO 2019.202 voor overzicht VO; BTO 2020.044, BTO 2021.008). De afgelopen jaren hebben, vooral ook op het vlak van bioinfomatica, veel ontwikkelingen plaats gevonden voor het analyseren en interpreteren van metagenomics- en metaproteomics-data. Hier liggen kansen voor toepassing in het onderzoek. Voor metaproteomics is momenteel nog weinig kennis beschikbaar over mogelijke toepassing in drinkwaterbereiding en het distributiesysteem.

Met deze bureaustudie wordt verkend of en in hoeverre metagenomics en metaproteomics toepasbaar zijn voor onderzoek in drinkwaterproductie en of daarmee waardevolle informatie over de microbiologische processen die plaatsvinden in zuivering en distributie wordt verkregen. Dit wordt gedaan met een literatuuronderzoek en door samenwerking met andere instituten. Op basis van deze inventarisatie zal worden vastgesteld of metagenomics en metaproteomics implementeerbaar zijn en of deze technieken waardevolle informatie geven over de microbiologische processen die plaatsvinden in de drinkwaterzuivering of distributie. Een concreet voorstel over vervolgonderzoek naar het ophelderen van microbiologische processen in de drinkwaterpraktijk maakt onderdeel uit van deze rapportage.



*Figuur 1. De verbinding tussen DNA, RNA en eiwitten en de verschillende analyses die hierop kunnen worden uitgevoerd: metagenomics (DNA-gebaseerd), metatranscriptomics (RNA-gebaseerd) en metaproteomics (eiwit-gebaseerd). Het figuur beschrijft het centrale dogma in de moleculaire biologie. Via **transcriptie** worden van specifieke genen op het DNA de RNA-moleculen gesynthetiseerd. Vervolgens worden deze RNA-moleculen gebruikt als template om eiwitten te maken in het **translatie**-proces. Bron: Bio-MENS.*

Metagenomics en metaproteomics zijn praktisch beter uitvoerbaar dan metatranscriptomics. Bij metatranscriptomics wordt gekeken naar het messenger RNA (mRNA) van een cel. Wanneer mRNA van specifieke genen aanwezig is, geeft dit aan dat deze genen actief worden getransleerd (Figuur 1). mRNA is hiermee de tussenstap tussen genen en eiwitten. Bij RNA-gebaseerd werk bestaan vooral uitdagingen rondom de stabiliteit van de RNA-moleculen. Concentratiestappen zorgen bijvoorbeeld voor sterke veranderingen in de samenstelling van het metatranscriptoom. Dergelijke beperkingen zijn veel minder van toepassing op metagenomics en metaproteomics, aangezien DNA en eiwitten veel stabieler zijn (DNA: meerdere dagen, eiwitten: enkele uren,RNA: enkele minuten). Eerder BTO-onderzoek heeft gekeken naar de toepassingsmogelijkheden van metatranscriptomics (BTO 2018.044). Metatranscriptomics is daarnaast aangestipt in de nieuwsbrief waar kansen van nieuwe moleculair-microbiologische methoden in de drinkwatersector van juni 2019 beschreven zijn (paragraaf 3.2 van deze nieuwsbrief). Hieruit volgde dat het uitvoeren van metatranscriptomics op drinkwatermonsters op dit moment nog veel haken en ogen kent, waardoor het lastig uitvoerbaar is. Limiterende factoren zijn voornamelijk de benodigde conservering van het RNA door zeer lage stabiliteit (met bijkomende transport- en veiligheidsissues) en de grote benodigde hoeveelheid RNA (5 µg), waarvoor in beide gevallen nog geen goede oplossingen voorhanden zijn.

1.1 Leeswijzer

In Hoofdstuk 2 wordt ingegaan op metagenomics als methode in het drinkwater-onderzoek. In 2.1 wordt ingegaan op de principes van metagenomics. In 2.2 worden de verschillende stappen in detail uitgewerkt met betrekking tot de laboratorium methoden, maar ook de data analyse (software en de principes achter deze stappen). In 2.3 wordt een overzicht gegeven van eerder uitgevoerde metagenomics-studies aan (drink)watermonsters. In 2.4 wordt een afweging gegeven van de meerwaarde en uitdagingen van metagenomics in het drinkwater-onderzoek. Ook wordt hier een overzicht gegeven van een metagenomics workflow die in het lab kan worden toegepast (Figuur 17). Tenslotte wordt in 2.5 een flowchart gegeven waarin voor de verschillende in het onderzoek relevante onderzoeksvragen en hypotheses de beste methode wordt weergegeven (Figuur 18).

In Hoofdstuk 2 wordt ingegaan op metaproteomics als methode in het drinkwater-onderzoek. In 3.1 wordt ingegaan op de principes van metaproteomics. In 3.2 worden de verschillende stappen in detail uitgewerkt met betrekking tot de laboratoriummethoden, software en de principes achter deze stappen. In 3.3 wordt een overzicht gegeven van eerder uitgevoerde metaproteomics-studies aan (drink)watermonsters. In 3.4 wordt een afweging gegeven van de meerwaarde en uitdagingen van metaproteomics in het drinkwater-onderzoek. Ook wordt hier een overzicht gegeven van een metaproteomics workflow die in het laboratorium kan worden toegepast (Figuur 26). Tenslotte wordt in 3.5 een flowchart gegeven waarin voor de verschillende in het onderzoek relevante onderzoeksvragen en hypotheses de beste methode wordt weergegeven (Figuur 27).

Vanwege het groot aantal begrippen dat binnen sequencing en sequencing-analyse wordt gebruikt is in dit rapport tevens een begrippenlijst opgenomen (Bijlage I; I.I: begrippenlijst metagenomics en I.II: begrippenlijst metaproteomics). Alle begrippen die in deze begrippenlijst zijn opgenomen zijn tijdens het eerste gebruik in de tekst vetgedrukt.

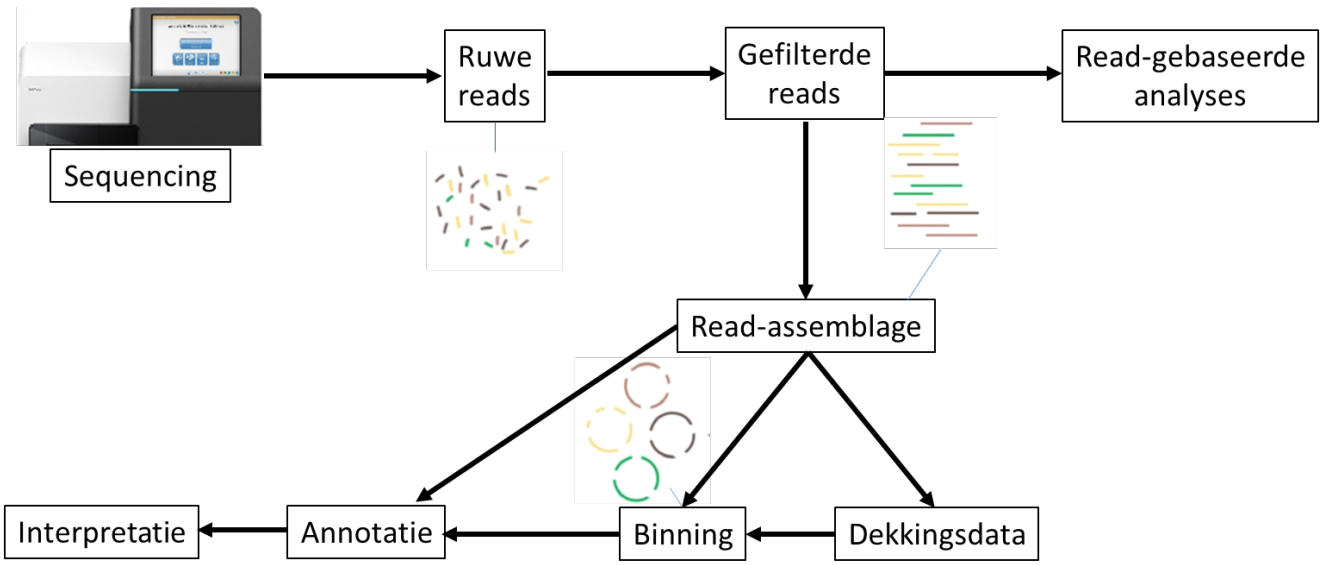
2 Het monitoren van de metabole potentie met metagenomics

2.1 De principes van metagenomics

Door de ontwikkeling van **next-generation sequencing** (**NGS**) platforms is het mogelijk geworden om snel veel DNA-data te genereren. Hierdoor is ook metagenomics sequencing zeer bereikbaar geworden. Het principe van metagenomics is dat alle DNA-fragmenten in een monster worden gesequenced en geanalyseerd. Dit in tegenstelling tot **amplicon sequencing** (hier is in onderdeel 2 van het BTO-project ingegaan) waarbij één enkel fragment van het **16S rRNA gen** wordt geamplificeerd middels PCR en gesequenced en geanalyseerd. Wanneer bij metagenomics voldoende wordt gesequenced kan hiermee in principe data van al het DNA-materiaal van alle soorten uit een specifiek milieu worden verzameld. Deze data kan vervolgens worden gebruikt voor een analyse van het metabole potentieel van het milieu, maar ook van de afzonderlijke soorten.
Een grote uitdaging bij metagenoom-analyses is dat diverse analysestappen noodzakelijk zijn om van ruwe DNA-data (de data die uit een sequencing-machine komt) naar data met een betekenis te gaan. In 2.2 wordt dieper ingegaan op de verschillende stappen in metagenoom-analyses. Ook is voor de uitvoering van deze stappen gedegen kennis van de bioinformatica en de resulterende data vereist. Ook hier wordt in 2.2 ingegaan.
Een limitatie van metagenomics is dat met deze methode alleen uitspraken kan worden gedaan over het metabole potentieel, ofwel de metabole processen die in theorie uitgevoerd kunnen worden. Er kan met metagenomics niets gezegd worden over de daadwerkelijke activiteit. Hiervoor zijn aanvullende technieken, zoals metaproteomics noodzakelijk. Hier wordt op ingegaan in hoofdstuk 3 van deze rapportage.

2.2 De stappen in een metagenoom-analyse

Binnen metagenomics sequencing vinden een aantal stappen plaats om van ruwe DNA sequencing-data naar data over de functionele potentie van verschillende micro-organismen te gaan (Figuur 2). Voor deze verschillende stappen is een zeer uitgebreid scala aan bioinformatische softwaretools beschikbaar. Hieronder wordt ingegaan op de verschillende stappen, de principes, methoden en uitdagingen. Ook worden per stap een aantal bioinformatische softwaretools uitgelicht die geschikt zijn voor toepassing in het onderzoek.
In het overzicht zijn de stappen en moleculaire procedures voor de sequencing niet meegenomen.



Figuur 2. Overzicht van de hoofdstappen in metagenoom data-analyse.
Overzicht van de verschillende stappen met als voorbeeld de Illumina MiSeq als sequencing platform. De ruwe reads betreft de data zoals deze door het sequencing-platform wordt aangeleverd. De gefilterde reads betreffen de reads die door de eerste kwaliteitscontrole komen. Vervolgens kunnen op deze gefilterde reads direct read-gebaseerde analyses worden gedaan. Ook kunnen deze reads worden gebruikt om langere DNA-sequenties te genereren in een proces dat read-assemblage (of 'assembly') genoemd wordt. Op de geassembleerde data kan vervolgens annotatie en interpretatie van de data worden gedaan om het functionele potentieel van een gehele microbiële dataset in kaart te brengen. Ook kan voor het reconstrueren van microbiële genomes gekozen worden. Met behulp van de dekkingsdata (dit is de informatie over hoeveel reads er gemiddeld gebruikt zijn om een assemblage te maken) kan vervolgens abundantie-informatie verkregen worden. Deze data kan later gebruikt worden om uitspraken over de abundantie van verschillende micro-organismen te doen. Wanneer nodig kunnen vervolgens in het binning-proces genomen van individuele micro-organismen worden gereconstrueerd. Om tot slot een betekenis aan deze genomes te geven worden de bins van individuele micro-organismen geannoteerd. Annotatie is het proces waarbij genen worden geïdentificeerd. Hiermee kan het metabole potentieel van specifieke gereconstrueerde genomes in kaart worden gebracht.

2.2.1 DNA-extractie

De DNA-extractie is de eerste stap in een metagenoom-sequencing protocol waarbij met moleculair biologische methoden het DNA selectief uit een monster geïsoleerd wordt. In het lab van KWR worden geoptimaliseerde protocollen gebruikt voor de extractie van DNA uit drinkwatermonsters. Hier wordt daarom in deze bureaustudie niet verder op ingegaan. Wel wordt hieronder kort ingegaan op belangrijke afwegingen en uitdagingen rondom DNA-extracties voor metagenoom-analyses.

Het is belangrijk om de **extractiebias** van een DNA-extractiekit in het achterhoofd te houden. Vaak wordt er daarom in metagenoom-analyses voor gekozen om twee verschillende DNA-extractiemethoden te combineren. Hierdoor wordt de extractiebias verlaagd. Aanvullend geeft het gebruik van twee verschillende extractiemethoden een extra dimensie voor de metagenoom binning-stap. Hier wordt in 2.2.7 ingegaan.
Anders dan bij PCR-gebaseerde Next Generation Sequencing (NGS)-analyses van bijvoorbeeld het **16S rRNA gen**, vindt bij metageoom-analyses geen **selectieve PCR**-stap op één specifiek gen plaats. Bij metagenoom-analyses wordt namelijk het gehele DNA van een monster gesequenced. Dit houdt in dat zowel het bacterieel DNA, maar ook het **eukaryoot** DNA van mensen, dieren, planten en schimmels en protozoën wordt meegenomen (BTO 2020.044 Microbiologische compositie en potentiële metabole processen tijdens drinkwaterdistributie). Hierbij kunnen uitdagingen ontstaan. Eukaryote organismen hebben namelijk een groter genoom dan bacteriën. Ter vergelijking, een gemiddelde bacterie heeft een genoom met een grootte van 5 miljoen basenparen (5 Mbp); kreeftachtigen zoals *Asellus* hebben een genoom van enkele honderden Mbp en schimmels een genoomgrootte variërend van circa 10 tot 200 Mpb. Een aanvullende uitdaging is dat schimmels vaak uit clusters van cellen met meerdere genoomkopieën bestaan, waardoor de hoeveelheid schimmel-DNA ook bij een lage hoeveelheid schimmel-biomassa hoog kan zijn. Hierdoor is er, afhankelijk van de aantallen bacteriën en eukaryote micro-organismen een bias naar eukaryoot DNA, doordat de hoeveelheid eukaryoot DNA in deze gevallen hoger is.

In de uiteindelijke data-analyse kan een beperkte hoeveelheid bacteriële sequencing-data uitdagingen opleveren. Om voorafgaande aan een uitgebreide metagenoom-run te controleren of er voldoende bacterieel DNA aanwezig is, kunnen algemene qPCR studies uitgevoerd worden op het 16S rRNA gen (als maat voor hoeveelheid bacterieel DNA) en het 18S rRNA gen (als maat voor de hoeveelheid eukaryoot DNA). Wanneer blijkt dat er te weinig bacterieel DNA aanwezig is, kan er voor gekozen worden om specifieke filtering van eukaryoot materiaal vóór DNA-extractie te doen. Dit kan bijvoorbeeld met filtratie op maat waarbij een filtergrootte gekozen worden waar prokaryote cellen wel door kunnen en eukaryote cellen niet. Deze methode maakt gebruik van het principe dat prokaryote cellen gemiddeld genomen kleiner zijn dan eukaryote cellen (0,1-5,0 µm versus 10-100 µm).

2.2.2 Sequencing

De keuze van de sequencing- technologie (platform) bepaalt de kwaliteit en de lengte van de individuele reads (Tabel 1). De twee belangrijke factoren bij metagenoom sequencing zijn de gemiddelde foutmarge van de sequencing-methode en de gemiddelde read-lengte. In de tabel is te zien dat Illumina sequencing een veel lagere foutmarge heeft dan de PacBio en MinION methoden. Echter, voor metagenomics sequencing vormen lange reads een groot voordeel voor zowel read-gebaseerde analyses (zie 2.2.3) en read-assemblage (zie 2.2.4). In de praktijk wordt daarom ook regelmatig gekozen voor de combinatie van short-read sequecing en long-read sequencing-technologieën. Hier wordt bij de relevante tussenstappen ook op ingegaan.

Tabel 1. Overzicht van de drie meestgebruikte sequencing-platforms, de gemiddelde read-lengtes en de gemiddelde foutmarges.

Sequencing platform	Read-lengte	Gemiddelde foutmarge ('error rate')
Illumina MiSeq	600 basenparen (2 x 300, paired-end protocol)	<0.1%
PacBio RSII	10-60 duizend basenparen	13%*
Oxford Nanopore MinION	>10.000 – 1 miljoen basenparen	10-15%**

**Hierbij is geen rekening gehouden met de mogelijkheid om de duurdere HiFi reads te verzamelen.*
***Recente ontwikkelingen bij Oxford Nanopore zorgen ervoor dat de foutmarge flink verlaagd wordt. Hier is in hoofdstuk 1 Metagenomics op ingegaan.*

2.2.3 Kwaliteitsfiltering

De eerste stap na sequencing is de kwaliteitsfiltering van de ruwe reads. Hierbij wordt gebruik gemaakt van kwaliteitsscores die door het sequencing-platform aan de output-data worden gegeven (de quality score, of Q-score). Vanuit de sequencing krijgt elke base van een individuele read een unieke Q-score (Tabel 4). Deze Q-score is een voorspelling van de waarschijnlijkheid van een fout in de identificatie van de betreffende base (het zogeheten 'base calling'). Deze Q-scores kunnen vervolgens worden gebruikt als filtering voor (delen van) de sequencing-data. In de kwaliteitsfiltering worden gehele reads met een gemiddeld te lage Q-score verwijderd.
De drempelwaarde hiervoor kan door de gebruiker worden vastgesteld. De keuze voor een drempelwaarde van de Q-score hangt grotendeels af van het sequencing-platform. Zo kan voor Illumina sequencing-data een Q-score van Q20 of Q30 prima gebruikt worden, terwijl voor standaard PacBio sequencing een Q-score van 10 hoog is.
Ook worden tijdens de kwaliteitsfiltering de reads die als te kort zijn verwijderd. Dit beslaan bijvoorbeeld reads die in het proces niet volledig gesequenced zijn.
Daarnaast kan er voor gekozen worden om uiteinden van reads met vaak een lage(re) Q-score af te knippen. Dit

proces heet **trimming**. Bij trimmen wordt van tevoren een drempelwaarde (**cutoff**) bepaald of wordt gekozen om een vaste lengte (aantal basenparen van de DNA-sequentie) van het begin en eind van de read af te knippen. Op basis van deze cutoff-waarde of de gekozen lengte worden de uiteinden van de reads verwijderd. Vervolgens kan, indien gewenst, opnieuw bepaald worden of de getrimde reads aan de minimale lengte-eisen voldoen.
Het doel van de kwaliteitsfiltering is om een dataset met reads van voldoende kwaliteit over te houden, zodat sequencing-fouten een zo klein mogelijk effect op de vervolg-analyses hebben. Een standaard softwaretool die gebruikt wordt voor deze stap is BBDuk. Deze tool is ontwikkeld door het DOE Joint Genome Institute en maakt gebruikt van k-mers voor de opzuivering van de read-data (BBDuk Guide - DOE Joint Genome Institute). Op het principe achter *k*-mers wordt in 2.2.6 dieper ingegaan.

Tabel 2. Overzicht van de kwaliteitsscores (Q-scores) die worden gebruikt voor sequencing-data.

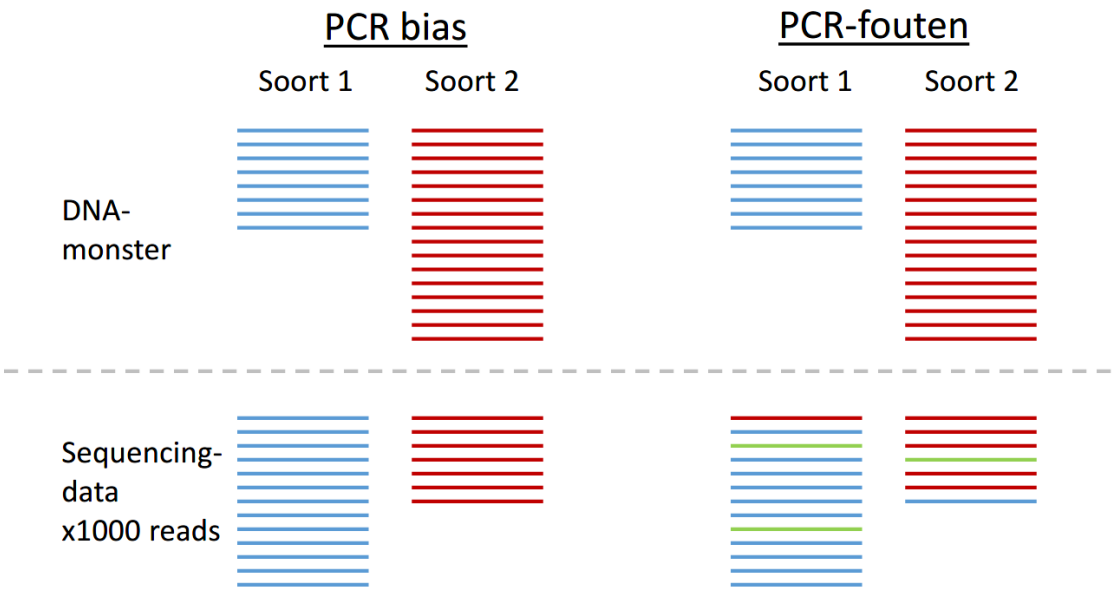
Kwaliteitsscore (Q-score)	Accuraatheid van de base calling
10	90%
20	99%
30	99,9%
40	99,99%
50	99,999%

2.2.4 Read-gebaseerde analyses

Op de kwaliteitsgefilterde reads kunnen direct read-gebaseerde analyses worden gedaan. Read-gebaseerde analyses kunnen antwoorden geven op de volgende vragen:
1. Wat is de globale **taxonomische compositie** van een monster?
2. Wat is het globale **metabole potentieel** van een monster?
3. Wat zijn verschillen tussen monsters wat betreft de globale taxonomische compositie en/of het globale metabole potentieel?
4. Welke specifieke **functionele genen** zitten in een monster?

Voor de analyse van de grove taxonomische compositie van een monster worden in de regel de 16S rRNA gen-bevattende reads uit een metagenoom-dataset gefilterd en geanalyseerd. Dit wordt gedaan omdat voor het 16S rRNA gen goede **referentiedatabases** beschikbaar zijn. Het proces is daarmee vergelijkbaar met amplicon sequencing, vaak NGS genoemd, aangezien in beide gevallen korte 16S rRNA gen-fragmenten worden gebruikt voor de analyse. Het belangrijke verschil is dat bij amplicon sequencing of NGS eenzelfde kort fragment van het 16S rRNA gen van verschillende soorten wordt geanalyseerd, zoals bijvoorbeeld het geval is bij de analyse van de V4-regio (zie de rapportage van activiteit 2 voor details). Bij metagenoom-gebaseerde analyse beslaan de reads echter *at random* het gehele genoom van micro-organismen en dus ook het gehele 16S rRNA gen van ongeveer 1500 basenparen. In principe is het daardoor ook mogelijk om uit metagenoom-data volledige 16S rRNA genen te reconstrueren. Hier wordt onder 2.2.5 kort ingegaan.
Het grootste voordeel van 16S rRNA gen-reads uit metagenoom-data is dat deze zonder 16S rRNA gen-gebaseerde PCR en de bijbehorende biases worden gegenereerd. Het is van belang om hierbij helder in beeld te hebben wat het verschil is tussen een PCR bias en fouten in de PCR (Figuur 3). Door de PCR bias geeft 16S rRNA gen-data uit metagenoom-datasets in principe een accurater en kwantitatiever beeld van de taxonomische samenstelling van

een monster. Echter geldt voor metagenoom-data dat de sequencing-diepte vaak beperkend is. Hierdoor zijn zeldzame groepen van micro-organismen vaak lastig te detecteren.



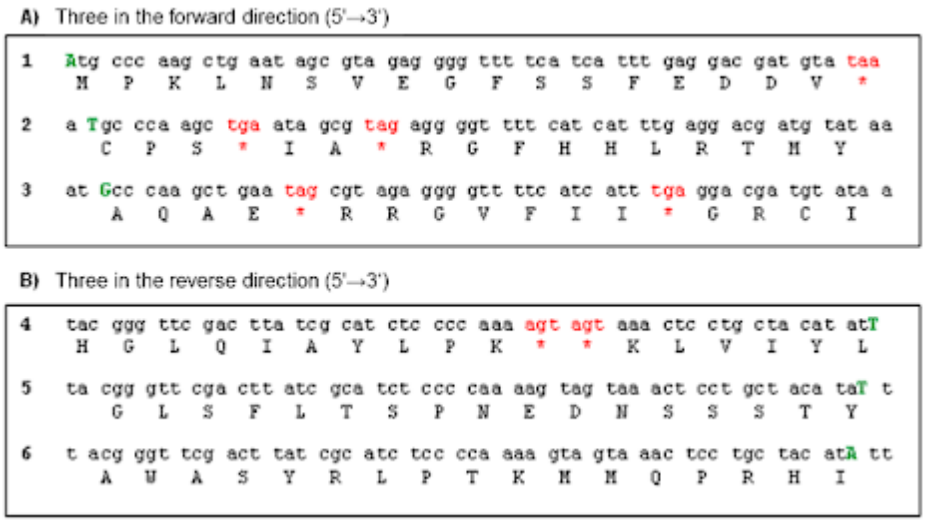
Figuur 3. Visualisatie van het verschil tussen de PCR bias en PCR-fouten. In dit voorbeeld is uitgegaan van een 1000x amplificatie van de DNA templates in de PCR-reactie. De bias wordt veroorzaakt doordat de PCR primers beter in staat zijn om DNA-fragmenten van de ene soort te amplificeren dan van de andere soort. Zoals zichtbaar is in de figuur bevat het DNA-monster verschillende aantallen van het DNA van soort 1 en soort 2. Door de PCR bias is in het uiteindelijk aantal na PCR verkregen fragmenten geen verband met het origineel aanwezige aantal DNA-fragmenten. PCR fouten ontstaan door onjuistheden tijdens de replicatie van het DNA in de PCR-reactie. Deze fouten zijn in principe random verdeeld. In de figuur is dit gevisualiseerd door PCR-fragmenten van een andere kleur in de eindreactie.

Wanneer het doel is om het algemene metabole potentieel van een microbiële gemeenschap in kaart te brengen, wordt een analyse van alle metagenoom-reads toegepast. Hierbij is de **resolutie** hoog, aangezien er veel reads en dus veel data beschikbaar zijn, omdat er een laag aantal monsters per sequencing run worden meegenomen. Een bijkomende uitdaging is wel dat de classificatie van deze reads over het algemeen niet heel nauwkeurig is, aangezien de databases voor functionele genen minder volledig en accuraat zijn. Wel kan met deze data een eerste, zeer grove uitspraak over het metabole potentieel gedaan worden. De informatie blijft slechts een indicatie, vooral door het feit dat de relatief korte reads slechts fragmenten van genen bevatten en dus een beperkte resolutie hebben.
Hieronder wordt ingegaan op Kaiju en MG-RAST, twee software-tools die gebruikt kunnen worden voor algemene read-gebaseerde analyses.

Kaiju

Een voorbeeld van een tool die gebruikt wordt voor de classificatie van alle reads in een metagenoom-dataset is Kaiju. Voor de analyses is een Kaiju web server beschikbaar, waardoor de analyse ook zonder beschikbare rekencapaciteit kan worden uitgevoerd. Daarnaast kan Kaiju op een eigen computer of server worden geïnstalleerd, waarmee de analyse kan worden uitgevoerd. Voor het gebruik van de nieuwste databases en voor de snelle analyse van grote hoeveelheden data is dit aan te bevelen. Het nadeel is dat de laatste update van de Kaiju databases voor de web server heeft plaatsgevonden in 2017.
Kaiju werkt op basis van de vergelijking van **aminozuursequenties** van de metagenoom-data met een referentiedatabase. Deze aminozuursequenties zijn de codes voor de eiwitten. De reden hiervoor is dat de aminozuur-sequenties nauwkeuriger functioneel zijn te classificeren.
In de eerste stap van een Kaiju-analyse dient de nucleotide-data (De ACTG-sequentie van het DNA) van de metagenoom reads vertaald (‘getransleerd’) te worden naar de aminozuur-sequenties. Hierbij wordt rekening

gehouden met de zes mogelijke **leesframes** voor de vertaling van de trinucleotiden (de code van 3 opeenvolgende nucleotiden die codeert voor een specifiek aminozuur) naar de aminozuren (Figuur 3). Op basis hiervan wordt een meest waarschijnlijk leesframe gekozen. In het voorbeeld in Figuur 3 is bijvoorbeeld te zien dat de 'forward direction' reading frames 2 en 3 verschillende stopcodons die het einde van een eiwit markeren kort na elkaar hebben. Dit is daarmee waarschijnlijk geen juiste vertaling van nucleotiden naar aminozuren.
Hetzelfde principe wordt overigens toegepast bij de annotatie van de geassembleerde sequencing-data. Hier wordt in 2.2.8i op ingegaan.



Figuur 4. Overzicht van de zogeheten leesframes ('reading frames') voor de vertaling van trinucleotiden naar aminozuren. Het symbool "" geeft een stopcodon weer. De drie leesframes in de voorwaartse richting zijn gelabeld met 1, 2 en 3. Deze worden alternatief gelabeld als +1, +2 en +3. De drie leesframes in tegengestelde richting zijn gelabeld met 4, 5 en 6. Deze worden alternatief gelabeld als -1, -2 en -3.*

Op basis van de meest waarschijnlijke annotatie voert Kaiju een snelle classificatie met een referentiedatabase uit. Beschikbare referentiedatabases zijn '**RefSeq** Genomes', '**NCBI BLAST *nr***' en 'NCBI BLAST *nr* + euk' waarbij ook Eukaryoten zijn opgenomen. Alle drie de databases bevatten een hoge microbiële diversiteit en zijn daarmee geschikt voor brede classificatiedoeleinden. De snelle classificatie met Kaiju is vaak te doen tot het taxonomisch niveau van **orde** of **familie**. **Genus**- en **soort**identificatie is gezien de beperkte taxonomische resolutie van de korte reads niet haalbaar.

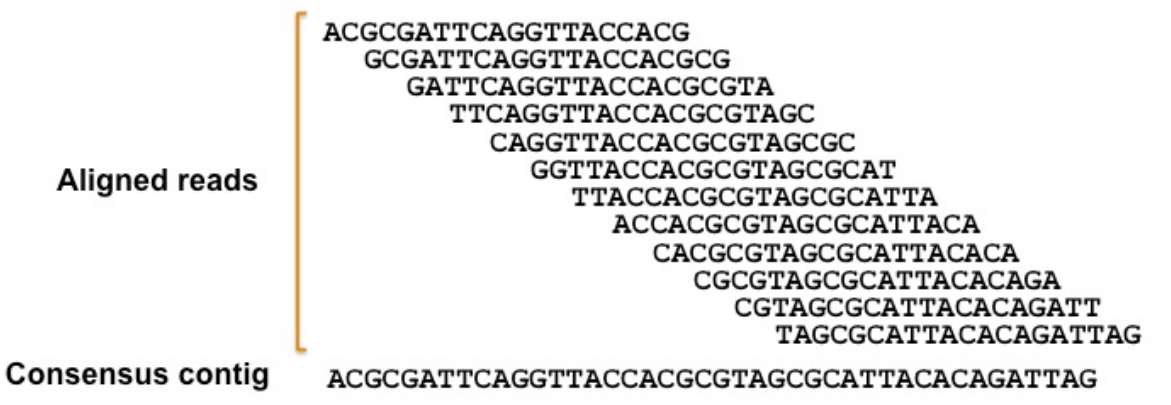
MG-RAST

Een tweede, veelgebruikte tool voor de analyse van ruwe sequencing reads is MG-RAST. De naam is een afkorting voor 'Metagenomic Rapid Annotations using Subsystems Technology'. Het is een online pipeline die automatisch functionele toewijzingen aan de reads doet, ofwel **annotatie** op basis van zowel nucleotide-gebaseerde als aminozuur-gebaseerde analyses die gebruik maken van referentiedatabases. Een nadeel is dat niet inzichtelijk wordt gemaakt welke referentiedatabases hiervoor worden gebruikt.
De analyse op de MG-RAST server duurt over het algemeen lang, vaak meer dan 1 maand. De analysetijd is mede afhankelijk van de openstelling van de te analyseren data. Wanneer wordt aangegeven dat de data na analyse openbaar beschikbaar mag worden gemaakt, krijgt de analyse een hoge prioriteit. Wanneer dit echter niet het geval is, krijgt de analyse een lage prioriteit en zijn wachttijden van meer dan 1 maand normaal. Dit is een van de grootste nadelen van de MG-RAST server.
Wel biedt de server uitgebreide opties voor de analyse van metagenoom-data. Zo bestaat de mogelijkheid om **fylogenetische** en functionele analyses op de data uit te voeren. Daarnaast kunnen ook verschillende metagenomes met elkaar vergeleken worden. Gedetailleerde informatie over de MG-RAST pipeline is te vinden op de MG-RAST - Wikipedia pagina.

Read-gebaseerde analyses die onder andere met Kaiju en MG-RAST kunnen worden uitgevoerd, geven een eerste inzicht in de grove taxonomische samenstelling en het functionele potentieel van een systeem. De genen coderen namelijk voor eiwitten met een bepaalde functie. Echter, het potentieel is niet direct gekoppeld aan specifieke microbiële soorten. Om de link tussen functies en soorten te kunnen leggen, dient read-assemblage, het genereren van dekkingsdata en binning te worden uitgevoerd (Figuur 2). Hier wordt onder 2.2.5, 2.2.6 en 2.2.7 dieper ingegaan.

2.2.5 Read-assemblage of 'assembly'

Het doel van read-assemblage, ofwel 'assembly', is om van korte(re) DNA-reads zo lang mogelijke sequenties te maken (Figuur 5). Deze langere DNA-sequenties worden contigs genoemd, wat staat voor 'contiguous DNA sequences' ofwel aaneengesloten DNA-sequenties. Op deze contigs bevinden zich volledige genen en genclusters, waarmee informatie kan worden verkregen over de potentiële rol van de microbiële gemeenschap in het milieu (zie 3.2.8). Het ultieme doel van read-assemblage is om volledige genomes van micro-organismen uit een monster te verkrijgen. In de praktijk geldt echter dat hiervoor niet voldoende sequencing-data aanwezig is, dat de beschikbare metagenoom-data te heterogeen is en dat ook de bioinformatische software-tools limitaties kennen. Voor volledige genoom-reconstructies is het daarom meestal noodzakelijk om aanvullende handmatige correcties en analyses te doen. Het genereren van een compleet microbieel genoom is daarmee een zeer tijdsintensieve en ambachtelijke klus. Wanneer het doel is om een algemeen beeld van het microbiële potentieel van een monster te krijgen, worden deze handmatige correcties niet uitgevoerd. Het is in dit geval namelijk voldoende als de genomes een relatief hoge compleetheid hebben (meestal ≥70% of ≥90%, afhankelijk van de onderzoeksvraag). Ook kan bij de vraag naar algemene diversiteits-analyses ook gebruik gemaakt worden van de geassembleerde reads zonder dat hier binning op is toegepast.
Voor de read-assemblage van DNA-data zijn een groot aantal (>150) tools beschikbaar. Het is door vaak kleine verschillen in de software lastig om hieruit de meest geschikte methode te kiezen. Hieronder wordt ingegaan op de read-assemblagetools die zich in de afgelopen jaren hebben bewezen als standaardmethoden voor de analyse van metagenoomdata. Deze tools vormen daarmee geschikte kandidaten om in toekomstig metagenoom-onderzoek te worden toegepast.

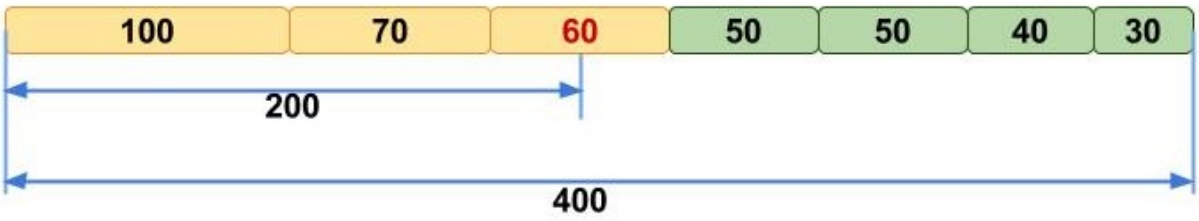


Figuur 5. Het basisprincipe van read-assemblage.
De figuur laat het basisprincipe van read-assemblage zien. Bovenaan zijn korte reads te zien die met elkaar overlappen en dus alignen. Door deze alignment kunnen de reads met assemblagesoftware in elkaar worden gepuzzeld in langere DNA-sequenties, contigs geheten (weergegeven als "Consensus contig"). Bron: newbler assembly

Assemblagesoftware

Zoals eerder vermeld wordt in de praktijk in de meeste gevallen Illumina sequencing gebruikt vanwege de relatief lage kosten voor een grote hoeveelheid sequencing-data met een lage foutmarge. De relatief korte reads vormen

echter de grootste uitdaging voor de read-assemblage. Grofweg geldt voor assemblage namelijk dat hoe korter de reads zijn, hoe meer reads er met minder sequencing-informatie aan elkaar moeten worden gepuzzeld (Figuur 5). Meestal worden voor de read-assemblage zogeheten *de novo* methode gebruikt. Bij deze methode wordt read-assemblage uitgevoerd zonder enige voorkennis en referentiedata. Voordelen van *de novo* methoden zijn dat de ontdekking en reconstructie van nieuwe genen en genclusters mogelijk is. Voor relatief minder goed gedocumenteerde milieus, zoals bijvoorbeeld drinkwatermilieus, verdienen *de novo* methoden daarom meestal de voorkeur. Wel is de novo assemblage vaak lastig met zeer korte reads. Dit heeft er mee te maken dat assemblage grotendeels afhankelijk is van overlap tussen de reads. Deze overlap is bij korte reads veel kleiner. Ook is een beperkte sequencing-diepte juist bij korte reads een extra uitdaging, omdat dan minder overlap tussen de reads gevonden kan worden.
Belangrijke praktische overwegingen voor de software-analyse zijn verder de benodigde tijd en het benodigde **RAM-geheugen**. Dit bepaalt namelijk of een bepaalde softwaretool met de huidige beschikbare **rekencapaciteit** van een computer of server zijn werk kan uitvoeren. In de studie van van der Walt en co-auteurs (2017) is een uitgebreide vergelijkingsstudie gedaan van verschillende assemblagetools op een negental metagenoom-datasets (bodems in Iowa, Oklahoma, permafrost-bodem, zee-ijsbloemen “frost flowers”, het Kolatameer, de Tara-oceaan en het darmmicrobioom van een kind, een Scandinavisch en Europees individu) en een drietal gesimuleerde datasets. Hieruit bleek dat SPAdes, metaSPAdes en MEGAHIT het beste presteerden (van der Walt et al. 2017) (Tabel 3). Hierbij dient te worden vermeld dat metaSPAdes de voor metagenoom-data geoptimaliseerde versie van SPAdes is. Het is daarom wenselijk om metaSPAdes te gebruiken, ondanks de soms op papier betere prestatiekenmerken van SPAdes op metagenoom-data.
In Tabel 3 is ook de CLC Genomics tool meegenomen. In vergelijking met de andere twee tools is de CLC genomics tool commercieel beschikbaar. Deze tool heeft een eenvoudige **grafische gebruikersomgeving** (‘graphical user interface’ of ‘GUI’) en is daarmee eenvoudiger te gebruiken zonder uitgebreide bioinformatica-achtergrond. Echter blijkt uit de vergelijking dat de methode weliswaar relatief snel is (analyse in <1 uur) en weinig rekencapaciteit gebruikt, maar dat de hoeveelheid contigs en zeker de hoeveelheid lange contigs (>1 kbp) vergeleken met MEGAHIT en metaSPAdes laag zijn (Tabel 3). Juist een hoge hoeveelheid langere contigs verhogen de resolutie van de geassembleerde data. Uit de vergelijkingsstudie is dus af te leiden dat MEGAHIT in verhouding zeer snel is en weinig rekencapaciteit nodig heeft.
Een veelgebruikte waarde voor de vergelijking van assemblage-tools is de **N50** (Figuur 4). De N50 is gedefinieerd als de lengte van de kortste contig op 50% van de totale metagenoom-lengte. Een N50 van 60 betekent bijvoorbeeld dat de helft van de contigs van de totale metagenoom-lengte korter is dan 60 basenparen, en de andere helft langer is dan 60 basenparen (Figuur 6).



Figuur 6. Visualisatie van een N50 van 60.
In dit voorbeeld heeft de N50 een waarde van 60. Te zien is dat de totale lengte van het metagenoom 400 basenparen is. Op 50% van de metagenoom-lengte, ofwel 200 zit een contig van 60 basenparen. De N50 is dus 60.

Ook uit een recente evaluatie van assemblage-software voor de reconstructie van antibioticaresistentie-genen volgt dat MEGAHIT en metaSPAdes een relatief laag aantal foutieve assemblages vertonen in vergelijking met andere tools, zoals bijvoorbeeld IDBA-UD (Brown et al. 2021). Een algemene uitdaging die in de evaluatie van Brown en co-auteurs (2021) naar voren kwam was de vorming van **chimere contigs**. Dit zijn contigs die zijn geassembleerd met reads die in werkelijkheid bij verschillende microbiële genomes horen. Dit is vooral een

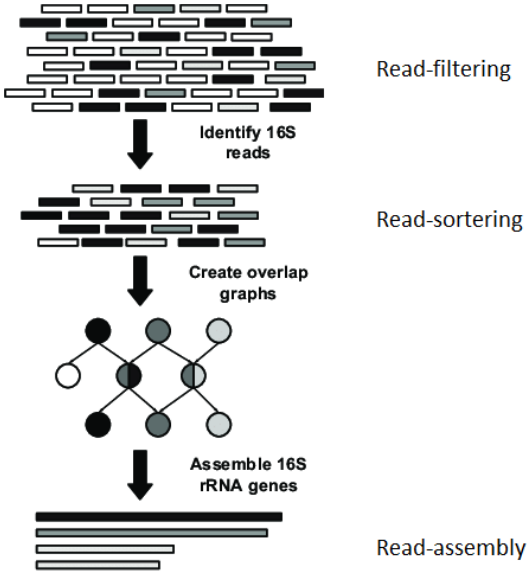
probleem met korte read-data, doordat korte regionen met dezelfde DNA-sequentie in twee verschillende soorten voor overlap kunnen zorgen.
Ook is de vorming van chimere contigs vaak een uitdaging wanneer in eenzelfde monster zeer op elkaar gelijkende soorten of stammen aanwezig zijn. Voor microbiële stammen geldt overigens in het algemeen dat huidige metagenoom assemblage-technieken vaak niet goed in staat zijn om deze van elkaar te onderscheiden, juist door de zeer op elkaar gelijkende genoom-sequenties. Het is daarom goed om in het achterhoofd te houden dat het onderscheid tussen microbiële stammen en varianten in een algemene metagenoom-analyse niet of nauwelijks mogelijk is.

Tabel 3. Overzicht van benodigde rekentijd en RAM-geheugen voor drie short-read sequence assemblage-tools, evenals het aantal contigs in de output-data, de N50 en het aantal lange contigs gelijk aan of groter dan duizend basenparen. Data overgenomen uit de analyse uitgevoerd door Brown et al. (2021).

Tool	Tijd in uren	RAM-geheugen in GB	Aantal contigs	N50	Lange contigs (≥1 kbp)
MEGAHIT	2.90	10.58	216.938	982	56.243
metaSPAdes	34.96	66.53	185.419	1124	48.640
CLC Genomics	0.98	16.23	50.716	880	10.720

16S rRNA gen-reconstructie

Het is met read-assemblage ook mogelijk om het 16S rRNA gen te reconstrueren en vervolgens aan de gegenereerde genomes toe te wijzen. Dit laatste is noodzakelijk om complete, gevalideerde metagenomes te reconstrueren. Het voordeel voor het gebruik van metagenoom-data voor 16S rRNA gen-analyses is dat deze data geen 16S rRNA gen-primerbias heeft. Echter, voor assemblagesoftware is de assemblage van 16S rRNA gen-data een grote uitdaging. Dit komt doordat het 16S rRNA gen een groot aantal geconserveerde regionen bevat die zeer vergelijkbaar zijn tussen soorten (zie Figuur 1 in de rapportage van onderdeel 2). Dit levert het hierboven genoemde probleem van chimere contigs op. Inderdaad is vaak te zien dat in standaard metagenoom assemblages de 16S rRNA gen-data verloren gaat, doordat de meeste 16S rRNA gen-reads van verschillende soorten in één enkele contig clusteren. Om specifieke 16S rRNA gen-gebaseerde read-assembly uit te voeren, zijn daarom veel handmatige aanpassingen en controles nodig. Alternatief kan gekozen worden voor gespecialiseerde software om de 16S rRNA gen-reads uit metagenoom-datasets te analyseren. Desondanks blijft reconstructie van het 16S rRNA gen vaak een zeer grote uitdaging met een beperkte kans van slagen.
Een voorbeeld van een assemblagetool voor de reconstructie van het 16S rRNA gen is REAGO (REconstruct 16S ribosomal RNA Genes from metagenOmic data) (Yuan et al. 2015). Deze softwaretool is gebaseerd op een eerste read-filtering om zo 16S rRNA-gen reads uit de metagenoom-data te halen (Figuur 7). Dit gebeurt door het vergelijken van alle reads met een 16S rRNA gen database, zoals de **SILVA database**. Vaak gebeurt dit met een kleinere referentiedatabase en relatief flexibele/'losse' instellingen, waardoor zo min mogelijk 16S rRNA gen-bevattende reads uit de metagenoom-data worden gemist. Vervolgens worden de reads softwarematig gesorteerd in clusters die één enkel 16S rRNA gen representeren (Figuur 7). Tot slot worden de reads geassembleerd en worden lange 16S rRNA gen-sequenties gevormd (Figuur 5). Hierbij houdt REAGO specifiek rekening met de secundaire structuur en geconserveerde regionen van het 16S rRNA gen. Met deze methode vindt een relatief nauwkeurigere reconstructie van gehele 16S rRNA-genen plaats. Op deze geassembleerde 16S rRNA-genen kunnen vervolgens analyses gedaan om microbiële soorten en de verwantschappen tussen deze soorten te analyseren.



Figuur 7. Overzicht van 16S rRNA gen-gebaseerde read assembly.

Combinatie short- en long-read technologieën in read assemblage

Vooral voor specifiekere doeleinden, zoals het compleet maken van individuele microbiële genomes, wordt vaak aanvullend long-read sequencing met PacBio of Oxford Nanopore toegepast. De long-read methoden dienen hierbij om gefragmenteerde genoom-puzzels, meestal gegenereerd met Illumina-data, in elkaar te zetten. Voor het assembleren van gecombineerde short-read/long-read datasets zijn specifieke tools als hybridSPAdes en OPERA-MS beschikbaar. OPERA-MS is gebaseerd op referentiedata. Dit maakt de methode beter geschikt voor goed gedocumenteerde milieus en minder voor bijvoorbeeld drinkwatermilieus waarvoor minder referentiedata beschikbaar is. hybridSPAdes is gebaseerd op de bewezen SPAdes algoritmes. De methode is gebaseerd op een initiële assembly van korte Illumina reads met SPAdes waarmee contigs worden gevormd. Hierna worden daarop de lange reads van PacBio of Oxford Nanopore sequencing gemapt. Door middel van de overlap kunnen vervolgens langere contigs gegenereerd worden. Deze methode maakt hierbij geen gebruik van referentiedata. Bij deze hybride contigs dient echter wel rekening gehouden te worden met het feit dat er regionen met hoge nauwkeurigheid zijn, opgebouwd uit een grote hoeveelheid Illumina reads, en regionen met lagere nauwkeurigheid, opgebouwd uit een kleine hoeveelheid lange reads met relatief hoge foutenmarge. Dit is vooral relevant voor de annotatie (2.2.8) en de interpretatie van de data (2.2.9) verderop in de metagenoom pipeline.

Uit het korte overzicht van de beschikbare assemblagesoftware is te concluderen dat MEGAHIT qua prestaties en benodigde analysetijd en rekencapaciteit (RAM-geheugen) de meest interessante open-source softwaretool is voor toepassing in metagenoom-analyses. De uitdaging voor MEGAHIT is dat gedegen kennis van de command line alsmede van het programma en de te kiezen parameters vereist is.
Als alternatief kan gekozen worden voor CLC Genomics, een gelicentieerd programma waarmee tevens assemblages uitgevoerd kunnen worden. Alhoewel voor het gebruik geen uitgebreide bioinformatica-kennis is vereist, volgt uit bovenstaande informatie wel dat de kwaliteit van de assemblage-data lager is.

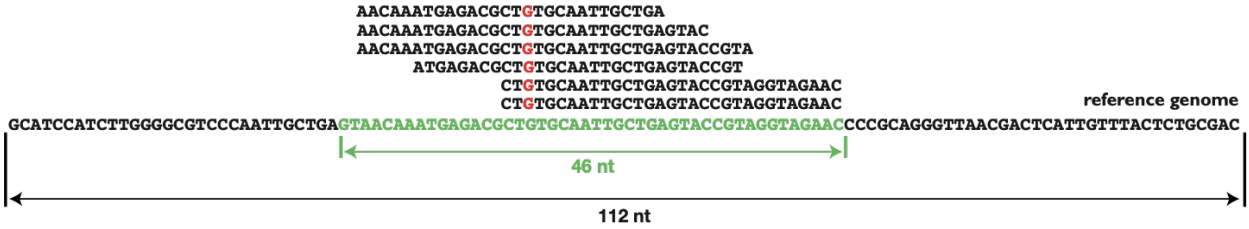
2.2.6 Dekkingsgraad of ‘coverage’

De dekkingsgraad of ‘coverage’ dient als basis om de abundantie van verschillende micro-organismen in een monster te bepalen. De dekkingsdata wordt verkregen op basis van de kwaliteitsgefilterde reads en de contigs. De dekkingsdata wordt gebruikt om antwoord te geven op de volgende vraag:
Wat is de relatieve abundantie van specifieke micro-organismen in een monster?

Om de dekkingsgraad te berekenen, worden de reads teruggeplakt op de contig die in de vorige stap is gevormd uit deze reads. Vervolgens wordt dan per base gekeken door hoeveel nucleotiden uit de reads deze gedekt wordt. Van het totaal kan vervolgens de **dekking/coverage** berekend worden. Voor deze berekening wordt de volgende formule toegepast:

Dekking/coverage = (aantal basen dat mapt op een contig) / (lengte contig)

Voor het terugplakken van de reads op de contigs wordt softwarematig een zogeheten '**alignment**' van de kwaliteitsgefilterde reads op de contigs uitgevoerd (Figuur 8).



Figuur 8. Overzicht van de procedure om dekkingsdata te verkrijgen.
Voor een contig wordt de dekkingsdata voor elke individuele nucleotide bepaald en vervolgens gemiddeld over de gehele contig. Zo is de coverage voor de in rood weergegeven 'G' bijvoorbeeld 5x en voor de meest rechts in groen voorkomende C 2x. Bron: ecSeq Bioinformatics.

Dekkingsdata kan verkregen worden met CLC Genomics of met verschillende open-source softwaretools zoals de breed toegepaste Burrows-Wheeler Alignment (BWA) tool (Li and Durbin 2009). De uitgevoerde berekeningen zijn zeer gestandaardiseerd, waardoor het in de praktijk weinig uitmaakt welke methode hiervoor wordt geselecteerd. De dekkingsdata is een zeer belangrijke dimensie voor metagenoom binning (2.2.7). Door de combinatie van dekkingsdata en binning kan de gemiddelde **abundantie** van micro-organismen verkregen worden. Omdat deze dekkingsdata een veel lagere bias heeft dan PCR-gebaseerde 16S rRNA gen-analyses, geeft deze data een veel beter benadering van de daadwerkelijke abundantie van de verschillende micro-organismen in het monster. Echter dient ook hierbij te worden vermeld dat de data door verschillende biases in onder andere de DNA-extractie maar zeker ook de verschillende stappen in de sequencing slechts een benadering is. Daarom moet ook in dit geval ook van relatieve abundantie worden gesproken.

2.2.7 Binning en het maken van bins / metagenome-assembled genomes (MAGs)

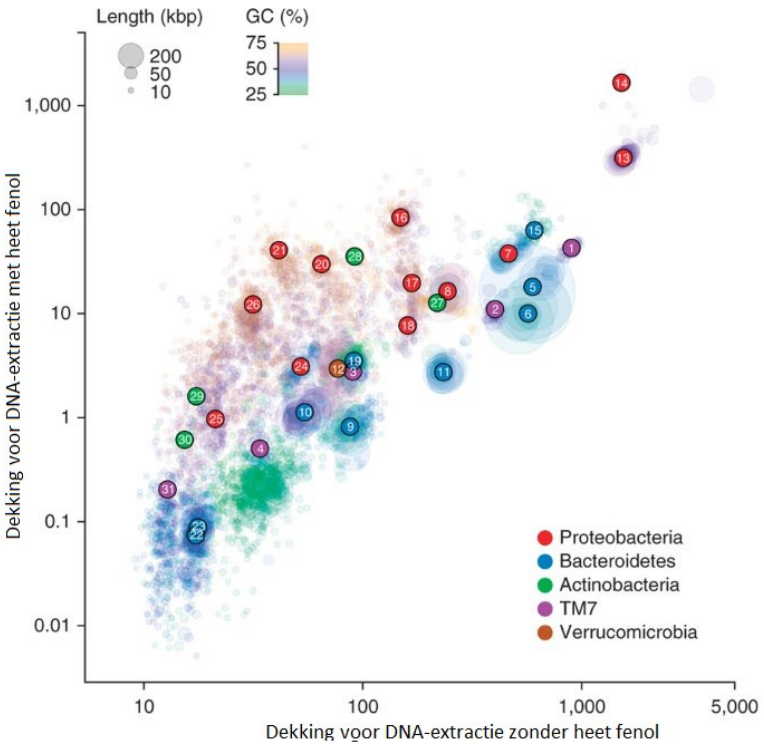
Binning is het proces waarbij uit de geassembleerde metagenoomdata individuele genomes worden gereconstrueerd. In de praktijk wordt gesproken van metagenoom bins of wanneer deze zijn gecheckt voor compleetheid worden ze 'metagenome-assembled genomes' of kortweg MAGs genoemd. Met individuele bins/MAGs kan veel soort-specifieke informatie verkregen worden.
De hoofdvragen die kunnen worden beantwoord, zijn:
1. Wat is de functionele en metabole capaciteit van specifieke micro-organismen in een monster?
2. Zijn er verschillen in soortsamenstelling tussen de verschillende monsters?

Tijdens het binningsproces worden de individuele geassembleerde contigs gegroepeerd in individuele clusters, bins geheten. Deze bins representeren (meestal) individuele soorten.
Binning gebeurt op basis van diverse kenmerken van de contigs die door read-assemblage zijn verkregen. Op een aantal belangrijke en veelgebruikte kenmerken wordt hieronder ingegaan.

Verschillende extractiemethoden

Zoals eerder aangegeven in 2.2.1 heeft elke DNA-extractiemethode een unieke bias. Dit betekent dat de DNA-samenstelling en dus ook de exacte microbiologische samenstelling tussen verschillende extractiemethoden anders is. Naast het feit dat het combineren van verschillende extractiemethoden de extractiebias verlaagd, zorgt dit ook bij binning voor een aanvullende dimensie die nuttig is. Wanneer op hetzelfde monster twee extractiemethoden worden toegepast, kunnen deze door de binning-software als twee verschillende data-inputs gebruikt worden. In aanvulling op de dekkingsdata, het GC-percentage en eventuele andere parameters vormt dit vaak een belangrijke parameter voor binning.

Een bekende studie waarin de effecten van verschillende DNA-extractiemethoden op binning is toegepast is de studie van Albertsen *et al.* uit 2013 waarbij een DNA-extractie met fenol en een extractie zonder fenol werd toegepast op een geactiveerd slibreactor ("activated sludge bed reactor") (Albertsen et al. 2013). Met deze methode konden uit de metagenoom-dataset enkele tientallen genomes worden gehaald omdat er met de ene DNA extractie methode meer DNA met hoger GC gehalte geëxtraheerd kon worden dan met de ander (Figuur 9).



Figuur 9. Voorbeeld van het effect van verschillende DNA-extractiemethoden op binning.
Voorbeeld van verschillen in dekking ("read coverage") voor twee verschillende DNA-extractiemethoden. In de eerste methode, weergegeven op de Y-as is bij de extractie heet fenol gebruikt. In de tweede methode, weergegeven op de X-as, is in de DNA-extractiemethode geen heet fenol gebruikt. De verschillen in DNA-extractie van verschillende organismen is terug te zien in de verschillen in dekkingsdata van de methode op de X- en Y-as. Figuur aangepast uit Albertsen et al. (2013).

Het principe zoals hierboven beschreven voor het gebruik van verschillende extractiemethoden geldt ook voor het gebruik van de hieronder genoemde factoren.

Dekkingsdata ("coverage data")

Verschillende micro-organismen hebben een verschillende abundantie in een monster. Deze abundantie werkt door in de sequencing-data. Grofweg geldt dat een micro-organisme met een hoge abundantie in een monster meestal ook een hoge abundantie in de sequencing-data heeft. Dit is terug te zien in de dekkingsdata. De dekkingsdata kan dus, zoals ook aangegeven in 2.2.5, worden gebruikt als parameter voor metagenoom binning.

Het GC-percentage

Het **GC-percentage** is vaak uniek voor groepen micro-organismen en zelfs individuele soorten en geeft daarmee een dimensie om contigs te categoriseren. Het GC-percentage is eenvoudig te berekenen met de volgende formule:

GC-percentage = Som(G+C)/Som(A+T+G+C) x 100%

Het GC-percentage vormt een belangrijke standaard dimensie tijdens binning.

Tetranucleotide frequenties

Een tetranucleotide is een reeks van vier nucleotiden achter elkaar (bijvoorbeeld 'ACTG'). Uit eerder onderzoek blijkt dat het patroon van verschillende tetranucleotiden uniek is voor verschillende micro-organismen (Teeling et al. 2004). Daarmee kan het patroon van tetranucleotiden gebruikt worden als een barcode van verschillende micro-organismen. Tetranucleotide-frequenties vormen tegenwoordig een belangrijke dimensie voor metagenoom-binning.

k-mers (k-meren)

***k*-mers** zijn stukken DNA-sequentie met een lengte van *k* nucleotiden. Waar bij **tetranucleotide frequenties** specifiek naar korte stukken van 4 nucleotiden wordt gekeken (dit is dus ook wel een 4-mer), gaat het bij *k*-mers meestal om stukken van enkele tientallen tot meer dan 100 nucleotiden. *K*-mers staan tegenwoordig aan de basis van de meeste binning-algoritmes. Het voordeel in vergelijking met tetranucleotiden is dat *k*-mers een hoger onderscheidend vermogen hebben. Het voordeel van *k*-mers is dat de berekeningen veel minder rekencapaciteit vereist dan wanneer gehele contig-sequences zouden worden vergeleken.

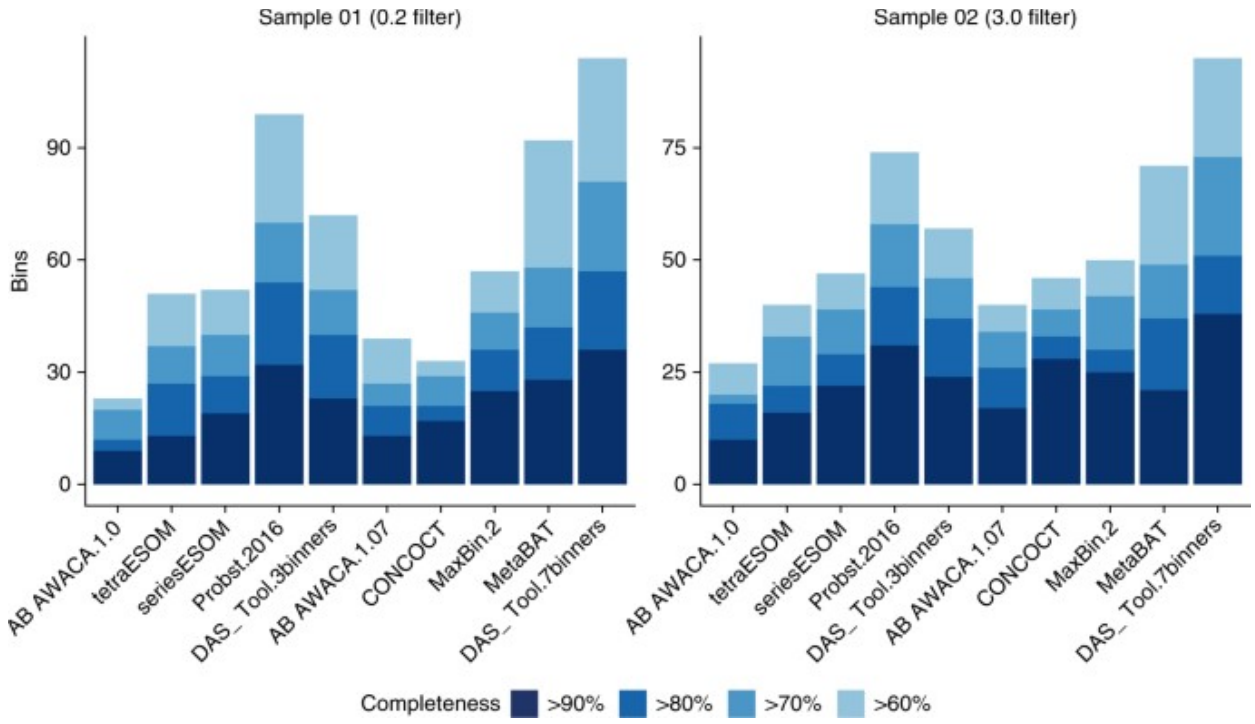
Referentiedata (uit databases)

De bovenstaande parameters worden onafhankelijk van referentiedata uitgevoerd. Echter kan ook juist referentiedata gebruikt worden voor de binning van contigs. Bij deze benadering wordt met behulp van referentiedatabases de waarschijnlijke identiteit van contigs vastgesteld. Op basis van deze waarschijnlijke identiteit kunnen de contigs vervolgens worden ingedeeld in verschillende bins. Aangezien dit zeer afhankelijk is van de kwaliteit van de contig en de kwaliteit van de referentiedata, heeft deze methode echter vaak een beperkte nauwkeurigheid en resolutie. Deze methode wordt dan ook niet standaard toegepast. Zeker voor relatief onbekende milieus met beperkte referentiedata, zoals drinkwatermilieus, is deze methode minder geschikt.

Er is een groot aantal dimensies die gebruikt kunnen worden voor metagenoom binning. De uitdaging bij veel huidige binning tools blijft echter dat vaak relatief weinig MAGs kunnen worden gereconstrueerd. Tevens zijn veel van deze MAGs relatief incompleet (≤50%) en van lage kwaliteit. Dit vormt een grote limitatie voor de analyses die op deze MAGs kunnen worden uitgevoerd. Wanneer het probleem zit bij de hoeveelheid en kwaliteit van de data is hier weinig aan te doen. Wanneer er voldoende data aanwezig is, kan met speciale "**wrapper tools**" een poging gedaan worden om betere MAGs te krijgen. Wrapper tools maken gebruik van meerdere binning-tools en selecteren na een vergelijking van de resultaten uit deze verschillende tools de beste MAGs. Met deze methode wordt over het algemeen de kwaliteit van MAGs verbeterd.

DAS Tool

DAS Tool is een zogeheten 'wrapper tool' die gebruik maakt van meerdere binning-algoritmes en uiteindelijk de beste resultaten van de verschillende algoritmes verzamelt. Dit proces wordt ook wel consensus binning genoemd. Het gebruik van DAS Tool resulteert in het algemeen in een groter aantal MAGs van gemiddeld hogere compleetheid (Sieber et al. 2018) (Figuur 10). In Figuur 10 is hiervan een voorbeeld weergegeven (Sieber et al. 2018). Hieruit valt af te leiden dat met het gebruik van DAS Tool een hoger aantal MAGs van gemiddeld een hogere compleetheid kan worden verkregen.
De compleetheid en kwaliteit van MAGs kan in detail bekeken worden met de tools CheckM en GTDB-Tk. Op beide tools wordt hieronder kort ingegaan.



Figuur 10. Gereconstrueerde genomes van de Crystal Geyser met verschillende binning tools.
De Crystal Geyser is een koudwatergeiser. Binning is uitgevoerd met 7 verschillende binning tools en met Das Tool als consensus binning tool op: AB AWACA.1.0, tetraESOM en seriesESOM (Das_Tool.3binners) en op deze 3 binners en AB AWACA.1.07, CONCOCT, MaxBin.2 en MetaBAT (Das_Tool.7binners). Probst.2016 beschrijft een manuele combinatie van de drie daarvoor genoemde tools, zoals eerder beschreven in Probst et al. (2016, Environ Microbiol). Met kleurcodering is de compleetheid van de MAGs weergegeven. Deze compleetheid is bepaald door een combinatie van de data uit DAS Tool en CheckM. Figuur overgenomen uit Sieber et al. (2018).

CheckM

CheckM is een tool waarmee de **compleetheid** (completeness), **contaminatie** (contamination) en de **heterogeniteit** (strain heterogeneity) van MAGs kunnen worden vastgesteld (Parks et al. 2015). Hiervoor zitten in de database van CheckM in totaal 5656 referentiegenomes, waarvan 2052 complete genomes en 3604 incomplete genomes. Voor de vergelijking zijn 43 geconserveerde markergenen geselecteerd die in het genoom van elk micro-organisme slechts één keer aanwezig zijn. Op basis van een scoring van aan- en afwezigheid kan vervolgens de compleetheid worden bepaald. Wanneer bijvoorbeeld alle 43 genen volledig aanwezig zijn, is een MAG volgens CheckM 100% compleet. Het kan echter voorkomen dat een markergen meerdere keren in een MAG voorkomt. Dit geeft aan dat er DNA-sequenties van meer dan één soort in deze MAG zitten. Dit geeft CheckM weer als contaminatie en heterogeniteit. Voor contaminatie geldt de definitie dat de verschillende versies van hetzelfde markergen een onderlinge identiteit hebben van <90%. Voor heterogeniteit geldt dat beide markergenen een identiteit hebben van >90%. Zo kan met de verschillende percentages voor compleetheid, contaminatie en heterogeniteit een uitspraak worden gedaan over de kwaliteit van een MAG.
Voor de kwaliteitsbeoordeling van MAGs zijn in 2017 standaarden gepubliceerd (Bowers et al. 2017).

Tabel 4. Kwaliteitscriteria voor de compleetheid van MAGs zoals vastgesteld door Bowers et al. (2017).

Kwaliteit MAG	Criteria
Finished MAG	Compleet genoom met een consensus errormarge van Q50 (≥99,999%) of hoger.
High-quality MAG	>90% compleet; <5% contaminatie
Medium-quality MAG	≥50% compleet; <10% contaminatie
Low-quality MAG	<50% compleet, <10% contaminatie

GTDB-Tk

GTDB-Tk is een tool die ontwikkeld is voor de classificatie van MAGs met de Genome Taxonomy Database (GTDB) (Chaumeil et al. 2019). In vergelijking met de CheckM database is de referentiedatabase van GTDB-TK gebaseerd op 194.600 genomes en daarmee een stuk completer. Met de GTDB-database wordt vooral voor Archaea een betere taxonomische classificatie gegeven, doordat deze in vergelijk met CheckM wel een goede referentiedataset heeft. Bij GTDB-Tk wordt de classificatie van genomes gebaseerd op de **RED-waarde**, een waarde die de relatieve evolutionaire afstand tussen genomes weergeeft en de average nucleotide identity (ANI, zie 2.2.9). De RED-waarde kan dus als vergelijkingsmateriaal worden gebruikt, waarbij geldt dat een hogere RED-waarde een grotere evolutionaire afstand aangeeft. Vaak wordt hierbij een schaal tussen de 0 en 1 gehanteerd waarbij 1 de hoogst aangetroffen evolutionaire afstand is tussen twee MAGs in de onderzochte dataset. Voor de classificatie met de GTDB-Tk wordt een MAG taxonomisch toegewezen aan de referentiedataset. Op basis hiervan geeft de GTDB-Tk de MAG een taxonomische locatie met als referentie het meest dichtbij staande referentiegenoom uit de database. GTDB-Tk biedt een beter alternatief voor de identificatie van MAGs. De compleetheid en contaminatie van MAGs wordt echter niet met deze software tool bepaald. Hiervoor is CheckM de standaard.

2.2.8 Annotatie

Het principe van annotatie is het toekennen van de eiwit-identiteit aan de hand van aminozuur-sequenties die verkregen zijn door de vertaling van nucleotiden-sequenties (zie Figuur 1 en Figuur 4). Door middel van annotatie-tools wordt biologische informatie (bijvoorbeeld informatie over genen, rRNA gen-subunits en tRNA-moleculen) gekoppeld aan de DNA-data van bijvoorbeeld de MAGs of van de geassembleerde data. Voor annotatie is een veelvoud aan softwaretools beschikbaar. Een niet-compleet overzicht van meer dan 140 tools is in 2020 door de National Bioinformatics Infrastructure Sweden (NBIS) gepubliceerd op github.

Offline tools (command line)

Een van de meest breed toegepaste en geaccepteerde command line-gebaseerde annotatietools is Prokka (Seemann 2014). Prokka is een snelle annotatietool die specifiek is ontwikkeld voor de analyse van prokaryote genomes. Het is een zogeheten *ab initio* methode, wat inhoudt dat de eiwit-coderende potentie alleen op basis van het genoom zelf bepaald wordt. Ter vergelijking, de 'similarity-based' (op gelijkenis gebaseerde) methode maakt gebruik van referentie-genomes van nauw verwante soorten. *Ab initio* methoden verdienen vooral bij minder goed gekarakteriseerde milieus, zoals drinkwatermilieus, de voorkeur.
Afhankelijk van de gekozen database verloopt de annotatie van een enkele MAG met Prokka zeer snel (<15

minuten). Wanneer echter wordt gekozen voor de zeer volledige NCBI RefSeq database die in de nieuwere versies van Prokka is opgenomen, is meer rekencapaciteit nodig (tot enkele tientallen GBs) en neemt de annotatie zeker enkele uren in beslag. Deze methode biedt vanwege een veel nauwkeurigere annotatie voor relatief onbekende soorten de voorkeur.
Een aanvullend voordeel van Prokka is dat tijdens de analyse direct bestanden worden gegenereerd voor aanvullende analyses en visualisatie in bijvoorbeeld de KEGG Automated Annotation Server (KAAS), in Artemis en voor het uploaden naar de Genbank database. Op deze tools wordt hieronder ingegaan.
Allereerst wordt ingegaan op een aantal online beschikbare tools voor de annotatie van MAGs.

Online tools

Er zijn verschillende online tools beschikbaar voor de annotatie van MAGs. Het voordeel van online tools is dat voor de data-analyse geen server-infrastructuur nodig is. Het nadeel van online tools is dat de sequencing-data dient te worden geüpload en dat de analyse van de data vaak langer duurt.
Door het NCBI wordt de Prokaryotic Genomes Automatic Annotation Pipeline via de email aangeboden. De analyse met deze pipeline duurt enkele dagen. Ook RAST is een webgebaseerde server waarmee zowel **bacteriële** en **archaeale** genomes kunnen worden geanalyseerd in vaak minder dan één dag (Aziz et al. 2008). Het verschil met MG-RAST is dat in RAST alleen genoom-data wordt verwerkt, terwijl met MG-RAST hele metagenoom-datasets worden geanalyseerd. Hierdoor is RAST in verhouding veel sneller dan MG-RAST. De tool xBASE2 is vergelijkbaar met RAST en geeft resultaten binnen enkele uren (Chaudhuri et al. 2007). De GO FEAT tool werkt niet op contig-sequenties en is daarmee niet toepasbaar op MAG-data (Araujo et al. 2018). Ook geldt dat de data die gesubmit is publiekelijk beschikbaar komt, wat binnen onderzoeksprojecten niet altijd wenselijk is.
Een alternatieve online tool is MicroScope (Vallenet et al. 2009). Bij deze tool wordt de data voor iedereen toegankelijk gemaakt na publicatie van het project. Ook is een voordeel van het platform dat veel gebruikers handmatig annotaties aanpassen, waardoor de MicroScope database nauwkeuriger wordt. Met 6.000 gesubmitte genomes en 2.700 actieve gebruikers (data uit 2017) wordt de tool echter beperkt toegepast (Vallenet et al. 2017). Wel dient hieraan te worden toegevoegd dat ook nu nog maandelijks nieuwe versies met aanvullende functionaliteiten worden uitgebracht. Zie voor meer informatie het MicroScope blog. Het kan daarom zinvol zijn de MicroScope tool bij metagenoom-analyses mee te nemen om de prestaties te onderzoeken.
De grootste beperking van de online methoden is voornamelijk dat hoge throughput niet mogelijk is. Dit heeft te maken met de hoge benodigde rekencapaciteit voor de annotatie van meerdere genomes. Aangezien uit een gemiddelde metagenoom-analyse zeker enkele tientallen MAGs komen, zijn de online methoden minder geschikt voor een brede annotatie. Wel zijn deze methoden zeer geschikt voor de analyse en vergelijking van enkele geselecteerde MAGs.

Annotatie-databases

Voor de analyse van gen-sequenties zijn een groot aantal publieke databases beschikbaar (Ejigu and Jung 2020). Grote publieke databases zijn GenBank, de European Nucleotide Archive (ENA) en de DNA Databank of Japan (DDBJ). Deze databases bevatten alle beschikbare informatie en zijn niet **manueel gecureerd**. Dit houdt in dat in de database ook onjuiste informatie voorkomt. Gezien het uitgebreide karakter vormen de databases van GenBank en ENA vaak de standaard voor annotaties.
Daarnaast is er ook de UniProt database, een combinatie van **UniProtKB/SwissProt** (ongeveer een half miljoen manueel samengestelde en gecureerde eiwit-sequenties) en UniProtKB/**TrEMBL** (180 miljoen automatisch geannoteerde sequenties op basis van de gecureerde database). Dit is een database van zeer hoge kwaliteit waarmee nauwkeurige identificatie van de functie mogelijk is. Een nadeel van deze database is dat de annotatie van relatief onbekende genen vaak niet mogelijk is. De UniProt database is daarmee meer geschikt voor de annotatie van specifieke MAGs en minder voor een grote diversiteit aan MAGs.
Ook is er de InterPro database waarmee informatie verkregen kan worden over **eiwitfamilies**, -**domeinen** (eiwitregionen die een eigen structuur hebben en zelfstandig stabiel zijn) en belangrijke gebieden zoals de **actieve site**, de locatie voor substraatbinding en geconserveerde regionen. Met InterPro kan dus meer gedetailleerdere

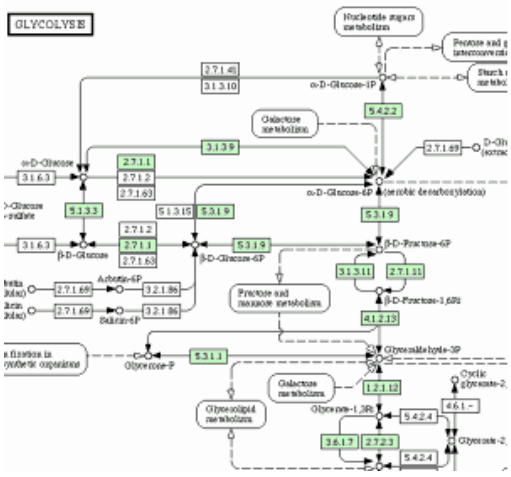
informatie over specifieke eiwitten verkregen worden door deze te vergelijken met goed in de database gedocumenteerde eiwitten. In het algemeen is de InterPro database minder geschikt voor brede annotatiedoeleinden, maar zeer geschikt om meer informatie over specifieke eiwitten te krijgen.
Daarnaast bestaan er ook databases voor de identificatie van zogeheten '**noncoding DNA**'. Dit beslaat bijvoorbeeld DNA dat voor niet eiwit-coderend **regulatoir RNA**, **pseudogenen** (kopieën van een functioneel gen die door veranderingen/mutaties hun functie hebben verloren) en **transposons** (DNA-fragmenten die hun positie in een genoom kunnen veranderen). Aangezien dit niet direct relevant is voor de analyse van het microbieel potentieel, wordt hier niet dieper op ingegaan.

Visualisatie van annotatiedata

Voor de interpretatie en eventuele manuele aanpassing van annotatiedata is visualisatie van de data een cruciaal hulpmiddel. Er bestaat een aantal visualisatietools die voor verschillende doeleinden wordt gebruikt. Hieronder wordt een tweetal veelgebruikte tools toegelicht.

De KEGG Automatic Annotation Server (KAAS)

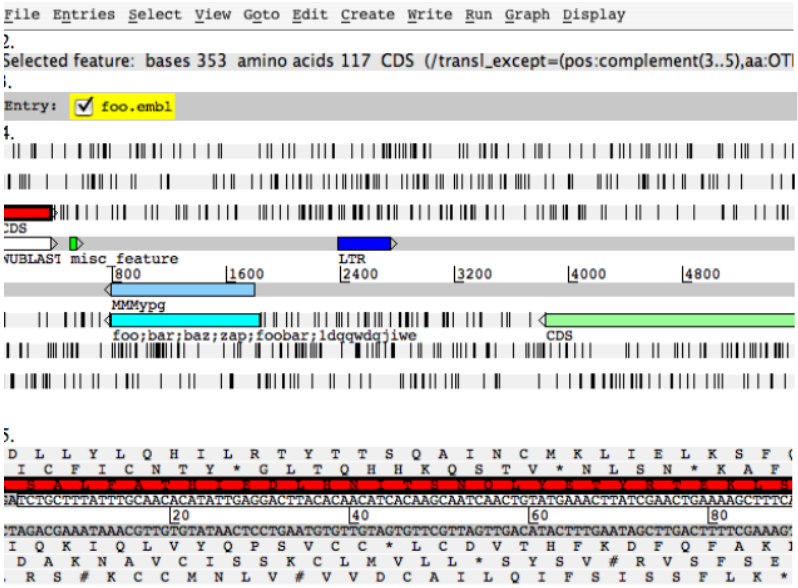
De online KAAS server wordt gebruikt voor het in beeld brengen van de functionele annotatie van MAGs op het niveau van specifieke **pathways**, zoals bijvoorbeeld pathways voor koolstoffixatie of de glycolyse (Moriya et al. 2007) (Figuur 11). De KAAS server maakt gebruik van annotatiebestanden die bijvoorbeeld door Prokka worden gegenereerd. Met de tool kan de compleetheid van verschillende pathways in kaart worden gebracht en kunnen eventuele missende genen in kaart worden gebracht.
Een uitdaging van deze methode is dat mis-annotatie leidt tot de onjuiste identificatie van genen die coderen voor enzymen in verschillende pathways, waardoor onjuiste aannames over het microbiële potentieel worden gedaan. Een pathway is een chemisch omzettingsproces dat gekatalyseerd wordt door vaak meerdere enzymen. Daarom is het belangrijk om voor de interpretatie van de data een zekere standaard voor de compleetheid van pathways wordt genomen. Een veelgekozen minimale compleetheid is hierbij 70%.
Naast misannotatie worden ook vaak delen van pathways gemist, doordat niet alle genen in de eerdere stappen geannoteerd kunnen worden. Ook kan het zijn dat zeer specifieke geannoteerde genen niet door KAAS worden erkend en dus in het overzicht ontbreken. Om dit op te lossen moeten de annotaties manueel gecheckt worden, wat een zeer tijdsintensieve klus is en voorkennis over microbiële fysiologie vereist. Daarnaast is er niet één of enkele tools die dit op kunnen lossen. Dit vraagt dus de hulp van veel tools die hier genoemd worden.



Figuur 11. Voorbeeld van een door KAAS gegenereerde pathway map voor de glycolyse.
De nummers zijn de Kegg Orthology identifiers (KO identifiers). Elk functioneel ortholoog gen (een gen dat voorkomt in verschillende soorten en eenzelfde biologische functie vervult) heeft eenzelfde KO identifier. Er geldt dus dat in de KAAS map specifiek naar functies gekeken wordt. De in het groen gemarkeerde KO identifiers zijn in de geanalyseerde MAG aangetroffen.

Artemis

Artemis is een gratis beschikbare tool van het Wellcome Sanger institute die gebruikt kan worden voor het visueel in kaart brengen van de geannoteerde genen en **ribosomale sequenties** op een genoom (Rutherford et al. 2000) (Figuur 12).
Artemis is een waardevolle tool om **genclusters** in kaart te brengen. Hiervoor is het echter van belang dat de MAG niet te gefragmenteerd is. Dit wil zeggen dat de MAG idealiter uit meerdere grote (>10 kb) contigs bestaat.
Daarnaast is Artemis een geschikte tool voor het opzoeken en extraheren van specifieke genen. Dit kan zowel in de nucleotiden-sequentie als in de aminozuursequentie. Vervolgens kan met deze sequenties aanvullend onderzoek worden gedaan.



Figuur 12. Voorbeeld van de visualisatie van een geannoteerd genoom in Artemis.
Met gekleurde blokken worden in het midden geannoteerde sequenties weergegeven. In het voorbeeld is de geannoteerde rode sequentie geselecteerd. Onderaan is van deze sequentie zowel de aminozuursequentie (onderaan in rood) en de nucleotidensequentie (direct daaronder) te zien. Via Artemis kunnen deze sequenties worden geëxporteerd. Bron: Artemis – Wellcome Sanger Institute

MetaCyc

MetaCyc is een gecureerde database van metabole pathways van alle domeinen (prokaryoten en eukaryoten). Momenteel bevat de MetaCyc database 2937 pathways van 3295 verschillende organismen. De database bevat data van pathways, **reactiemechanismen**, enzymen en genen voor zowel het **centraal energiemetabolisme** als van **secundaire metabole processen** in organismen. Het doel van de database is om van elk metabool proces in micro-organismen een experimenteel geverifieerd voorbeeld te hebben. De tool kan net als KAAS gebruikt worden voor het voorspellen van metabole pathways in gesequencete microbiële genomes.

Aanvullende analyses

Ook kan specifiek gekeken worden of de gevonden enzymen van organismen binnenin de cel (**intracellulair**), door de celmembraan (transcellulair) of buiten de cel (**extracellulair**) aanwezig zijn. Hiervoor kunnen tools als Protter en de TMHMM Server gebruikt worden. Deze tools kijken naar specifieke domeinen (aminozuur-sequenties) in de geannoteerde eiwitten om te zien of er **signalen** aanwezig zijn die aangeven dat een eiwit specifiek binnen of buiten de cel voorkomt.

Protter visualiseert eiwitvormen en maakt een interactieve visualisatie en integratie van geannoteerde eiwitten. Dit wordt gekoppeld aan experimentele datasets die in de Protter-database aanwezig zijn. Zo wordt gevisualiseerd waar in of buiten de cel een eiwit waarschijnlijk aanwezig is. De TMHMM Server zoekt daarbij naar transmembraan-helices in de eiwitten van een database. Als deze transmembraan-helices aanwezig zijn, dan kan worden aangenomen dat een eiwit van nature deels in het membraan zit.

Limitaties en uitdagingen bij annotaties

Het is essentieel dat het annotatieproces altijd handmatig wordt geïnspecteerd en waar nodig wordt bijgestuurd. Bijsturing houdt in de praktijk vaak in dat het annotatieproces opnieuw wordt doorlopen met minder strenge of juist strengere parameters/instellingen.

Daarnaast geldt dat een annotatie slechts zo goed is als de beschikbare database. Voor een groot aantal genen geldt dat er geen goede referentiedata beschikbaar is. Deze genen worden tijdens de annotatie gelabeld als '**hypothetical protein**', oftewel een hypothetisch eiwit waarvan de functie nog niet is vastgesteld. Over het algemeen geldt dat bij monsters die genomen zijn uit het milieu, zoals drinkwatermonsters, een groter percentage van de genen uiteindelijk als 'hypothetical protein' gelabeld zullen worden.

Inzicht verkrijgen in de mogelijke functie van 'hypothetical proteins' is zeer complex en arbeidsintensief. Het is bijvoorbeeld niet haalbaar om bij de analyse van een complex monster voor de verkregen MAGs aanvullend onderzoek naar deze 'hypothetical proteins' te gaan doen.

Aanvullend geldt dat voor de interpretatie en vergelijkbaarheid van de annotatie van verschillende MAGs uit verschillende studies het van belang is dat gestandaardiseerde annotatielabels gebruikt worden. Een gouden standaard binnen de annotatie van eiwitcoderende genen is het gebruik van Enzyme Commission (**EC**)-**nummers**. Een overzicht van de gebruikte EC-codering is te vinden op de website van de NC-IUBMB. EC-nummers worden als standaard gebruikt in annotaties met Prokka en dienen als standaardreferentie in KAAS pathway mapping (Figuur 11).

Samengevat kan met de annotatie van MAGs uit metagenoom-data dus een uitspraak gedaan worden over de aan- en afwezigheid van bepaalde enzymen in pathways die bepaalde metabole processen laten verlopen. De analyse richt zich op de genen die voor de enzymen coderen en geeft daarmee een breed beeld van het metabole potentieel van een MAG. Uiteraard kunnen aanvullende analyses op specifieke genen gedaan worden, indien deze genen in de MAG aanwezig zijn.

Aanvullend kunnen Hidden Markov Models (HMMs) worden gebruikt. **HMMs** worden veelvuldig toegepast binnen NGS-analyses. HMMs vormen een effectieve methode om niet-geannoteerde gen-sequenties met een bepaald motief te koppelen aan bekende genen met eenzelfde motief. Hierbij wordt standaard gebruik gemaakt van aminozuursequenties.

Anders dan bij algemene annotatietools gaat het hierbij specifiek om het identificeren van motieven die karakteristiek zijn voor bepaalde genen. Deze motieven worden door de HMM-tool in kaart gebracht op basis van referentiedata. De kracht van HMMs zit in de identificatie van genen die door algemene sequentie-vergelijking gemarkeerd worden als 'hypothetical protein', omdat de brede sequentie-vergelijking een onvoldoende gelijkenis ziet tussen een gen en de referentiedata.

Wanneer tijdens een standaard annotatie blijkt dat een zeer groot deel van de genen van de MAGs als 'hypothetical protein' wordt geannoteerd (dit kan in de praktijk rond de 70% liggen), kan er voor worden gekozen om een methode met HMMs toe te passen op de data. Prodigal is een veelgebruikte softwaretool waarin genen worden geïdentificeerd met HMMs (Hyatt et al. 2010).

2.2.9 Overige analyses

Er zijn verschillende aanvullende analyses die met metagenoomdata kunnen worden uitgevoerd. Hieronder wordt een aantal aanvullende methoden voor algemene MAG-analyses benoemd.

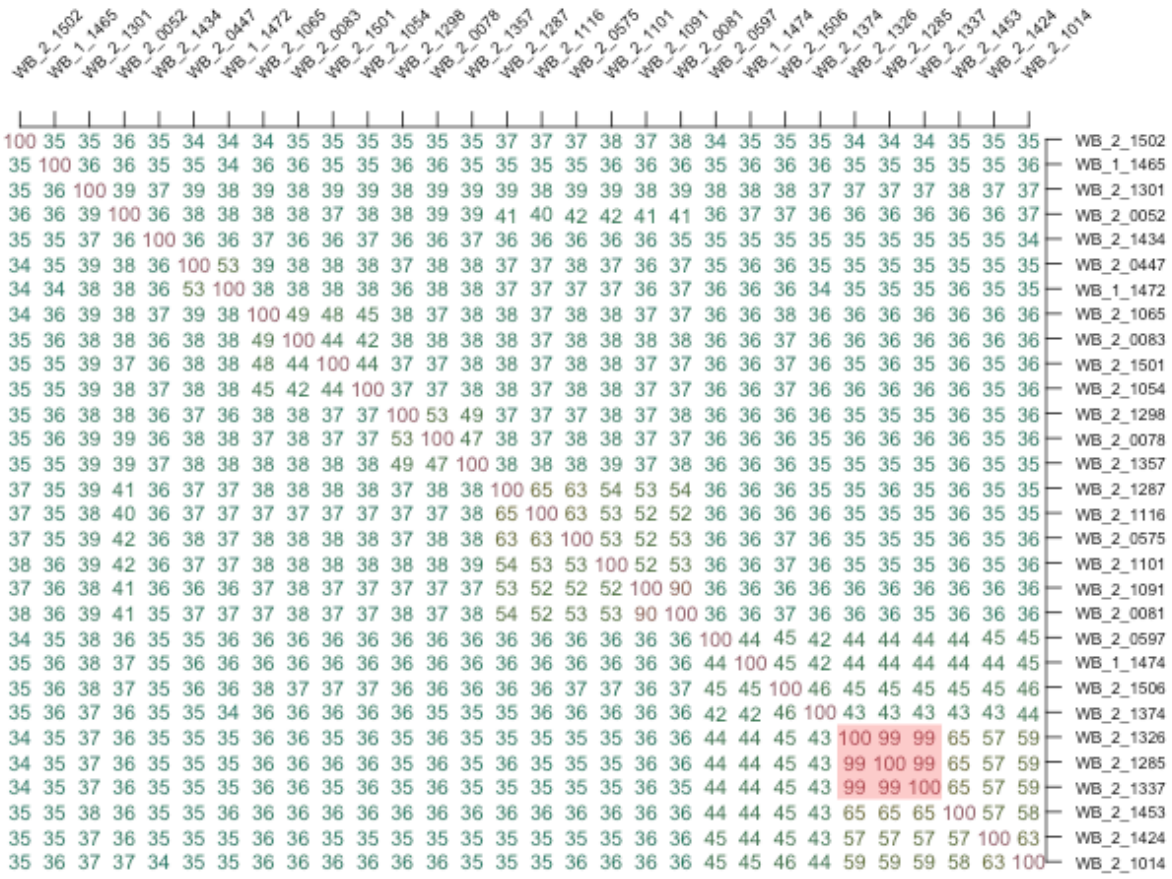
Genoomvergelijkingen

Indien het gewenst is om de identiteit en het metabole potentieel van verschillende microbiële MAGs met elkaar te vergelijken, kan gebruik worden gemaakt van verschillende genoomvergelijkingsmethoden. Hieronder wordt ingegaan op een aantal brede genoomvergelijkingsmethoden. Deze methoden kunnen worden toegepast op de verkregen MAGs.

Average nucleotide identity (ANI) en average amino acid identity (AAI)

Twee globale vergelijkingsmethoden zijn de average nucleotide identity (**ANI**) en de average amino acid identity

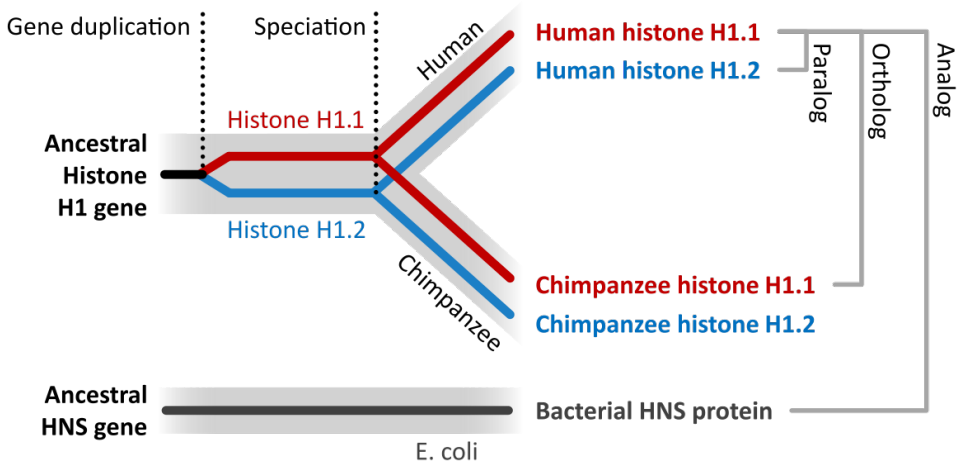
(AAI) (Konstantinidis and Tiedje 2005). De ANI en de AAI worden gebruikt als algemene parameters om de verschillen tussen twee of meerdere MAGs in kaart te brengen. Bij de ANI wordt gekeken op nucleotiden-niveau, wat een betere evolutionaire resolutie heeft. De mutaties in het DNA worden namelijk meegenomen in de analyse. Bij de AAI wordt gekeken naar aminozuur-sequenties, wat een betere functionele resolutie heeft. Zowel de ANI- als AAI-methode werkt met het maken van grote matrix-tabellen waarbij de percentuele verschillen tussen alle in de analyse toegevoegde MAGs worden berekend en weergegeven (Figuur 13).
Beide methoden geven een gemiddeld percentage van respectievelijk het aantal identieke nucleotiden en aminozuren in de gedeelde orthologe genen. Orthologe genen zijn genen die afstammen van éénzelfde voorouder-gen en in verschillende soorten organismen dezelfde biologische functie hebben.
ANI en AAI waarden geven een globaal beeld van de identiteit van verschillende MAGs. De waarden kunnen gebruikt worden om te bepalen of twee verschillende MAGs tot eenzelfde soort behoren. Hierbij wordt een cutoff-waarde van 95% voor ANI en 95-96% voor de AAI gehanteerd (Khaleque et al. 2020). De ANI- en AAI-waarden zijn eenvoudig te berekenen met de online ANI/AAI Tool van het Kostas Lab (Rodriguez-R and Konstantinidis 2016).



Figuur 13. Voorbeeld van een AAI matrix gegenereerd door de ANI/AAI-Matrix tool van het Kostas Lab.
Hierboven is een AAI matrix te zien waarin een 31-tal MAGs op aminozuursequentie met elkaar vergeleken zijn. Uit een dergelijke vergelijking kunnen genomes met een hoge AAI-identiteit worden geïdentificeerd. In deze matrix zijn deze genomes in rood gemarkeerd.

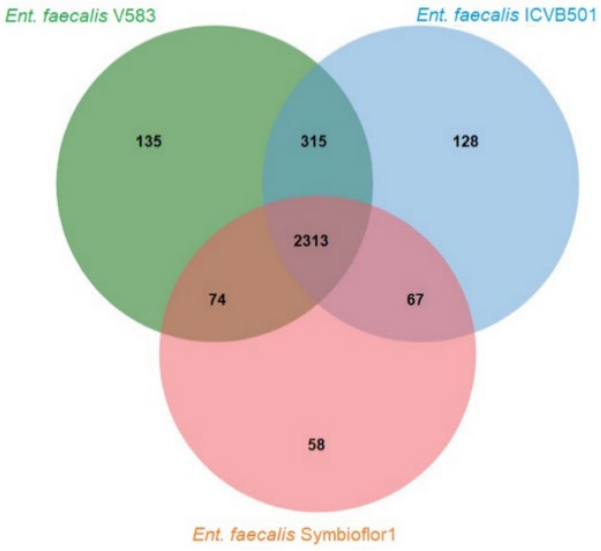
Orthologe en paraloge genen/genclusters

Genoomvergelijkingen kunnen ook op basis van het aantal orthologe en paraloge genen en genclusters worden uitgevoerd. Het principe achter orthologe genen is hierboven toegelicht. Paraloge genen zijn genen die ontstaan zijn door een duplicatie van één gen in het genoom. Hierna zijn deze paraloge genen geëvolueerd en coderen deze voor vergelijkbare, maar niet identieke functies.



Figuur 14. Het verschil tussen ortholoog, paraloge en analoge genen.
In dit overzicht is het voorbeeld van de histon H1 genen te zien. Het menselijk genoom bevat twee histon H1 genen: H1.1 en H1.2. H1.1 en H1.2 zijn paraloge genen van elkaar. Deze genen zijn ontstaan door een duplicatie van het originele H1 gen in het humane genoom. Hetzelfde is het geval in het genoom van de Chimpansee. De genen H1.1 en H1.2 tussen de Chimpansee en de mens zijn orthologen van elkaar. Ze hebben hetzelfde voorouder-gen (gen H1) en vervullen dezelfde functie. De humane histon-genen H1.1 en H1.2 zijn verder een analoog van het bacteriële HNS gen. Analoog wil zeggen dat beide genen dezelfde functie hebben, maar geen evolutionaire geschiedenis delen. Ze zijn dus apart van elkaar ontstaan. Bron: Wikipedia.org

De overlap tussen orthologe en paraloge genen en genclusters is een maat voor de functionele gelijkenis van verschillende MAGs. OrthoVenn2 is een webserver voor de genoomvergelijking van orthologe en paraloge genclusters van verschillende MAGs (Xu et al. 2019). In OrthoVenn2 worden venn-diagrammen gegenereerd waarin het aantal orthologe genclusters tussen verschillende MAGs wordt geïdentificeerd (Figuur 15). Het aantal paraloge genclusters wordt aanvullend gerapporteerd.

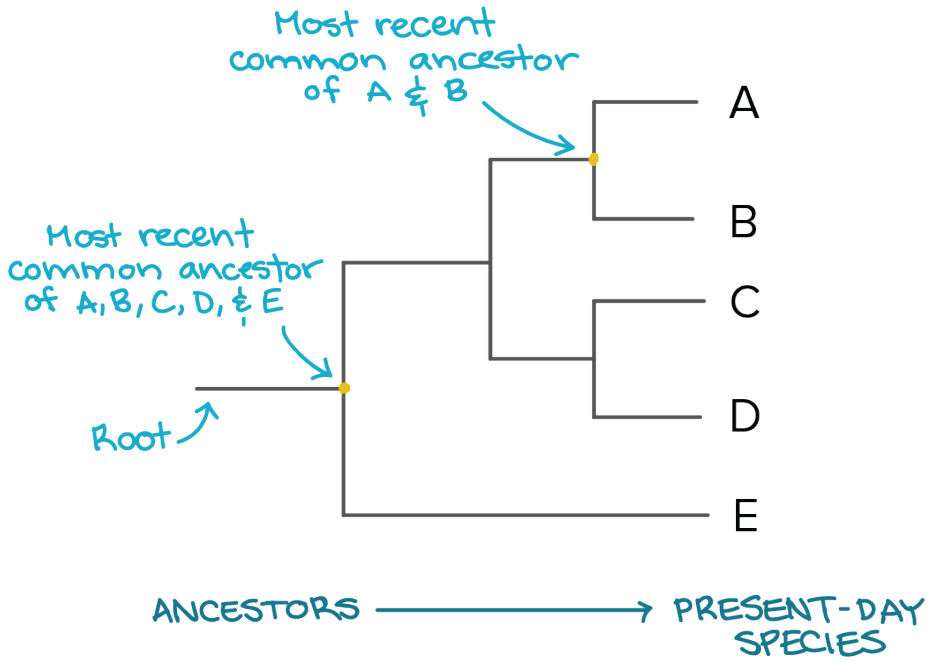


Figuur 15. Voorbeeld van een venndiagram gegenereerd met OrthoVenn2.
De getallen in het venn-diagram geven het aantal orthologe genclusters weer. Figuur overgenomen uit een studie van Eveno en co-auteurs (Eveno et al. 2021).

Fylogenetische bomen

Wanneer de vraag bestaat om het taxonomisch of evolutionair verschil tussen MAGs in kaart te brengen, kan hiervoor een **fylogenetische boom** worden gemaakt. Een voorbeeld met aanvullende uitleg is hieronder in Figuur 16 weergegeven.

Binnen tools als CheckM en GTDB-Tk bestaat de mogelijkheid om fylogenetische bomen te generen.
Ook kan aanvullend op basis van het 16S rRNA gen of geselecteerde functionele genen een fylogenetische boom gemaakt worden. Een breed toegepaste en gratis softwaretool voor het maken van fylogenetische bomen met zelf-geselecteerde genen is MEGA (Hall 2013).



Figuur 16. Voorbeeld van een fylogenetische boom.
Het beginpunt van een fylogenetische boom wordt de 'root' ofwel de wortel genoemd. De T-splitsingen geven het punt aan waar een gemeenschappelijke voorouder is afgesplitst in twee verschillende soorten, zoals bijvoorbeeld zichtbaar is bij de T-splitsing tussen A en B. De horizontale afstand is een maat voor de evolutionaire afstand tussen de voorouders (ancestors) en de huidige soorten (a, B, C, D en E). De horizontale afstand die tussen de soorten A, B, C, D en E moet worden doorlopen is een maat voor de evolutionaire afstand tussen de huidige verschillende soorten. Zo staan soort A en B relatief dicht bij elkaar (het horizontale pad tussen A en B is kort), terwijl soort A en E ver van elkaar af staan (het horizontale pad tussen A en E is lang). Voorbeeld overgenomen van Khan Academy (khanacademy.org).

De detection-based approach

Het is tevens mogelijk om een detectie-gebaseerde benadering ofwel '**detection-based approach**' te gebruiken voor de analyse van sequencing-data. Hierbij is het doel om met hoge precisie, maar relatief lage gevoeligheid de aanwezigheid van specifieke organismen vast te stellen. Vaak gaat het hier om pathogenen. Vragen die met de detectie-gebaseerde methode kunnen worden beantwoord, zijn:
1. Zijn specifieke micro-organismen (bijvoorbeeld pathogenen) aanwezig in het monster?
2. Zijn specifieke functionele genen (bijvoorbeeld voor antibioticaresistentie) aanwezig in het monster?

Voor specifieke vragen, zoals de hierboven genoemde vraag rondom aanwezige resistentiegenen voor antibiotica, zijn de detectie-gebaseerde methodeninteressant. Er bestaan verschillende tools waarmee dit kan worden uitgevoerd.
Taxonomer is een tool met een eigen web interface waarmee de antibioticaresistentie voor virussen, bacteriën en schimmels kan worden bekeken. De tool is toegespitst op humane pathogenen.
SURPI, hetgeen staat voor Sequence-Based Ultra-Rapid Pathogen Identification is een tool gericht op de identificatie van antibioticaresistentie in klinische monsters. De tool heeft daarmee dezelfde mogelijke beperkingen als Taxonomer. Desondanks bieden beide tools interessante mogelijkheden voor de identificatie van voor de menselijke gezondheid relevante resistentiegenen voor antibiotica.
In eerder BTO-onderzoek is tevens gekeken naar resistentiegenen voor antibiotica in oppervlaktewater- en drinkwaterzuvieringsprocessen (BTO 2017.042). Dit is destijds met 16S rRNA gen-sequencing voor kwantificatie en PCR op specifieke merkergenen van antibioticaresistentie gedaan. Een dergelijke benadering met de in deze studie

aangewezen genen kan ook in een metagenoom-studie worden uitgevoerd. Daarnaast kan het gebruik van een van de hierboven genoemde tools zorgen voor een bredere screening van de aanwezigheid van resistentiegenen. Eerdere verkennende metagenoom-gebaseerde onderzoeken zijn bij KWR uitgevoerd met de DeepARG tool (Arango-Argoty et al. 2018). DeepARG is ontwikkeld om resistentiegenen voor antibiotica in een metagenoom-dataset in kaart te brengen.

2.2.10 Geautomatiseerde analyses

De mogelijkheden voor volledig geautomatiseerde analyses van metagenoom-data zijn vooralsnog erg beperkt. Voor elke studie bestaan namelijk specifieke wensen en bestaat de noodzaak om handmatig de instellingen aan te passen en de parameters van de verschillende softwaretools te optimaliseren. Wel bestaan er een aantal tools waarin verschillende software-programma's worden geïntegreerd. Het bekendste en meest complete programma hiervoor is KBase. In KBase kan metagenoomdata worden geüpload en kunnen beschikbare programma's als MaxBin2, DAS Tool en CheckM worden gedraaid.
Voor het eenvoudig delen en visualiseren van grote hoeveelheden metagenoom-data is CyVerse interessant. Daarnaast kan ook IMG-ER gebruikt worden waarin de annotatie, analyse en visualisatie kan worden uitgevoerd. De mogelijkheden van IMG-ER zijn echter beperkt in vergelijking met de eerder genoemde programma's.
Voor de specifieke analyse van infectieziekten en pathogene micro-organismen is PATRIC interessant. Ook dit programma biedt tools voor de visualisatie van de data.
In het algemeen geldt echter dat de geautomatiseerde tools momenteel nog geen complete analyse van metagenoom-datasets mogelijk maken. Wel valt er een snelle ontwikkeling in deze tools te zien, waardoor het belangrijk is om de ontwikkelingen hierin nauwlettend in de gaten te blijven houden.

2.3 Eerdere metagenoom-studies aan (drink)water

Er is uitgebreid onderzoek gedaan aan watersystemen met metagenomics. Binnen de drinkwaterzuivering zijn diverse onderzoeken gedaan naar de microbiologie in verschillende stappen van de zuivering, naar effecten van desinfectiestrategieën, maar bijvoorbeeld ook specifiek naar resistentiegenen voor antibiotica in drinkwater.

Drinkwaterzuivering – In een onderzoek aan een drinkwaterzuiveringsinstallatie in de Pearl River Delta in China is de microbiële samenstelling tot familieniveau in kaart gebracht (Chao et al. 2013). In dit onderzoek is daarnaast gekeken naar verschillende biotische factoren, zoals de aanwezigheid van vetzuren, aminozuren maar ook het metabole potentieel (koolstoffixatie, stikstofmetabolisme) als verklarende factoren van de diversiteit van drinkwaterpopulaties. Hieruit kwam naar voren dat de microbiële gemeenschap significant verandert tijdens de drinkwaterzuivering en tevens bescherming ontwikkeld tegen eventuele chlorering.
Desinfectiestrategieën – In een onderzoek naar de effecten van verschillende desinfectiestrategieën (drinkwater behandeld met vrij chloor en chlooramines) is specifiek gekeken naar de aanwezigheid van *Mycobacterium*, *Legionella* en Cyanobacteriën als onderdeel van de bacteriële populatie (Gomez-Alvarez, Revetta, and Santo Domingo 2012). Voor de eukaryote populatie is specifiek naar de aanwezigheid van amoeben, ciliaten en slijmzwammen ('slime molds') gekeken. De toepassing van desinfectiemiddelen was duidelijk gelinkt aan virulentiefactoren en antibioticaresistentie-mechanismen in de ecologische netwerken.

Filtersystemen – Een onderzoek aan verschillende drinkwaterzuiveringsfilters heeft gekeken naar de microbiologie op verschillende filtersystemen, waaronder de snelle zandfilter, granulair actieve koolfilter en de langzame zandfilter (Oh, Hammes, and Liu 2018). Hieruit bleek dat vooral bacteriën die actief zijn in de stikstofcyclus verrijkt zijn op de zandfilters, hetgeen veroorzaakt wordt door de aanwezigheid van ammonium in het water. De microbiële samenstelling van zandfilters is ook in kaart gebracht door Poghosyan *et al.* met een focus op de stikstof- en

methaancyclus (Poghosyan et al. 2020). Bij een onderzoek bij pompstation Breehei (WML) is naast het in kaart brengen van de algemene microbiologische samenstelling daarbij vooral onderzoek gedaan naar de samenstelling en identiteit van de functionele genen om zo meer inzicht te krijgen in de verantwoordelijke micro-organismen binnen deze nutriëntencycli. Hieruit volgt dat de stikstof- en methaancyclus cruciaal zijn in deze zandfilters.
Eukaryote populatie – In een onderzoek aan een waterzuiveringsinstallatie in China is parallel 18S rRNA sequencing en metagenomics toegepast om specifiek de eukaryote diversiteit in verschillende zuiveringsstappen in kaart te brengen (Li et al. 2021). Hieruit bleek de laagste eukaryote diversiteit (soortenrijkheid) aangetroffen te worden in gedesinfecteerd water. Over de verschillende zuiveringsstappen werd over het algemeen een grote variatie in verschillende groepen (Arthropoda, Ciliophora, Ochrophyta, en Rotifera als dominante fyla). Uit dit onderzoek bleek dat vooral het metabole potentieel sterk verschilde tussen de zuiveringsstappen, al is niet bekend wat dit voor de daadwerkelijke uitgevoerde metabole processen betekent.
Antibioticaresitentie – Er is veel interesse voor antibioticaresistentie in drinkwater en de mogelijke gezondheidsrisico's die dit met zich meebrengt door de aanwezigheid van potentieel pathogene micro-organismen. Een onderzoek aan waterzuiveringssystemen voor thuisgebruik heeft de zogeheten resistoom in deze filtersystemen waarin actieve kool gebruikt wordt onderzocht (Zhou et al. 2021). De fysieke vorm van het actieve kool blijkt daarbij een groot effect op de samenstelling van de antibioticaresistentiegenen (ARGs) te hebben. Deze ARG kon in dit onderzoek ook aan specifieke micro-organismen gelinkt worden. In een ander onderzoek is het effect van verschillende desinfectiestappen op de antibioticaresistentie van de microbiële populatie in een drinkwaterzuiveringsinstallatie in Yancheng City in China zijn in kaart gebracht (Jia et al. 2020). Zuiveringsstappen zijn hier coagulatie, sedimentatie, zandfiltratie en desinfectie. Als desinfectie worden drie verschillende strategieën toegepast, te weten: antimicrobiele hars (antimicrobial resin, AR)/chloride (Cl_2), AR/UV en AR/ozon (O_3). Hierbij is gekeken naar een 80-tal subtypen van antibioticaresistentie bepaald aan de hand van ARGs-OAP, een online pipeline om antibioticaresistentiegenen in metagenoom-data te detecteren (Yang et al. 2016). Met een dergelijke studie kan op hoge resolutie in kaart worden gebracht welke resistentiegenen in de verschillende stappen aanwezig zijn. Ook binnen het BTO-onderzoek is naar ARGs gekeken. Uit een verkenning naar de toepassingsmogelijkheden van MinION sequencing bleek dit platform geschikt om ARGs te sequencen, al beperkt de relatief hoge foutenmarge van de sequencer momenteel nog de toepasbaarheid (Heijnen, 2019). Snelle ontwikkelingen in de technologie maken dat toepassingen binnen enkele jaren wel te verwachten zijn.

2.4 De meerwaarde en uitdagingen van metagenomics

Uit dit onderzoek kan geconcludeerd worden dat metagenomics door recente ontwikkelingen op het gebied van bioinformatica-tools waardevolle informatie kan geven over de microbiologische processen in drinkwaterzuivering en -distributie. Met metagenomics is het mogelijk om informatie te krijgen over de microbiële diversiteit tot op soortniveau en de metabole potentie (de biochemische processen) van een microbiële populatie. De technologie heeft daarmee een toegevoegde waarde op 16S rRNA gen-gebaseerde studies waarmee alleen de microbiële diversiteit in kaart wordt gebracht.
De ontwikkelingen in de bioinformatica-software maken de berekeningen sneller en efficiënter (er is minder rekencapaciteit nodig). Aanvullend vormen long-read technologieën, vooral de Oxford Nanopore MinION, op korte termijn waarschijnlijk een belangrijke bijdrage in de kwaliteitsverbetering van metagenoom-data.
De belangrijkste uitdagingen blijven enerzijds de benodigde rekencapaciteit voor metagenoom-datasets en anderzijds de benodigde bioinformatica-expertise van een groot aantal verschillende softwaretools. Voor de rekencapaciteit geldt, hoewel bioinformatica-tools efficiënter worden, dat de analyse van metagenoomdata geavanceerde serverruimte vereist.
De meerwaarden en uitdagingen van metagenomics zijn hieronder samengevat.

Meerwaarde

- Taxonomische informatie over de microbiële populatie tot op soort- en zelfs stam-niveau met het 16S rRNA gen en functionele genen.
- Informatie over het metabole potentieel van de microbiële populatie op basis van de genenpool.
- Nauwkeurigere informatie over de relatieve abundantie van de aanwezige micro-organismen (geen/nauwelijks* primer bias).
- Een bredere diversiteit van aanwezige micro-organismen kan worden gedetecteerd wanneer diep genoeg wordt gesequenced (geen/nauwelijks* primer bias)
- Metagenoom-data kan gebruikt worden voor de ontwikkeling van specifieke primers voor (q)PCR van 16S rRNA genen of functionele genen.
- Geschikt voor de analyse van monsters waarover zeer weinig informatie bekend is.

Uitdagingen

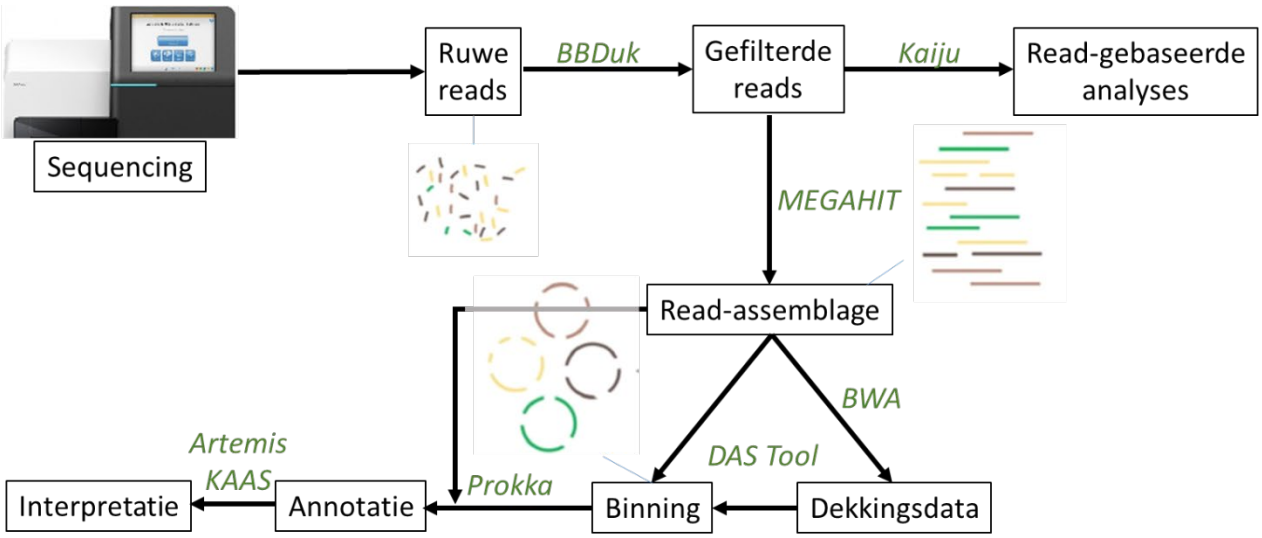
- Er is een groot aantal bioinformaticatools nodig voor de uitvoering van een metagenoom-analyse. Ook is voor elke stap een groot aantal verschillende tools beschikbaar.
- Software-analyse is tijdsintensief en vraagt om voldoende rekencapaciteit (data-server), bioinformatica-expertise en expertise over microbiële fysiologie en ecologie.
- Eukaryoot DNA kan een verstorend effect op de metagenoom-data hebben, waardoor er te weinig bacterieel DNA wordt gesequenced, en vice versa bij sequencing eukaryoten.
- Er is vaak beperkte informatie over laag-abundante micro-organismen (incomplete MAGs). Meer/dieper sequencen verhoogt de kosten en de rekencapaciteit en benodigde tijd voor verwerking van de data.
- De aanwezigheid van groot aantal ‘hypothetical proteins’ (genen waaraan tijdens annotatie geen potentiële functie kan worden toegekend) levert uitdagingen rondom de data-interpretatie op het vlak van microbiële fysiologie en ecologie.
- Minder geschikt voor de analyse van een groot aantal monsters (>10), door beperkingen in de rekencapaciteit.

*Ook voor het verhogen van de hoeveelheid te sequencen DNA wordt bijvoorbeeld bij Illumina sequencing een “algemene” PCR amplificatiestap gebruikt. Dit wordt gedaan met algemene primers die desondanks een beperkte bias kunnen hebben.

2.4.1 Implementatiemogelijkheden van metagenomics

De hoofdvraag van deze bureaustudie binnen het BTO-project Moleculair Microbiologische Methoden betreft de ontwikkeling van methoden waarmee de rol van microbiologie bij zuiveringsprocessen en in het distributiesysteem kan worden onderzocht. In dit hoofdstuk is daarbij specifiek naar de potentie van metagenomics gekeken.
In diverse projecten bij KWR zijn metagenomics-analyses uitgevoerd, maar hierbij bleek het verkrijgen en de verwerking van de data niet optimaal, waardoor de kracht van metagenomics niet volledig werd benut.
Als onderdeel van deze bureaustudie naar de implementatiemogelijkheden van metagenomics is een overzicht samengesteld voor de benodigde en meest geschikte softwaretools voor een standaard metagenoom-analyse (Figuur 17) waarbij de metagenomics-data zo veel mogelijk volledig wordt benut. Dit overzicht kan als basis

gebruikt worden voor een bioinformatica-pipeline voor de verwerking van metagenoom-data. In deze metagenomics-pipeline wordt zowel gekeken naar de microbiële identiteit (Kaiju, CheckM, GTDB-Tk) als naar het metabole potentieel dat in de genenpool aanwezig is (Prokka, en aanvullende visualisatie met KAAS en Artemis).



Figuur 17. Overzicht van softwaretools (in groen) voor de verschillende stappen in een metagenoom-analyse.
In dit overzicht zijn de softwaretools opgenomen die nodig zijn voor de analyse van de data. Voor de kwaliteitscontrole van de ruwe reads is dit BBDuk; voor de directe analyse op kwaliteitsgefilterde niet-geassembleerde reads is dit Kaiju; MEGAHIT wordt aanbevolen voor de assemblage van de reads; BWA is aanbevolen om de dekkingsdata te verkrijgen; met DAS Tool (en een minimum van 3 binning-tools binnen DAS Tool, waarbij CONCOCT MaaxBin2 en MetaBAT de standaard zijn) kan het beste binningsresultaat worden verkregen; Prokka is de meest gebruikte en meest breed inzetbare annotatietool. Visualisatie kan vervolgens uitgevoerd worden met KAAS en Artemis (niet weergegeven in deze figuur).

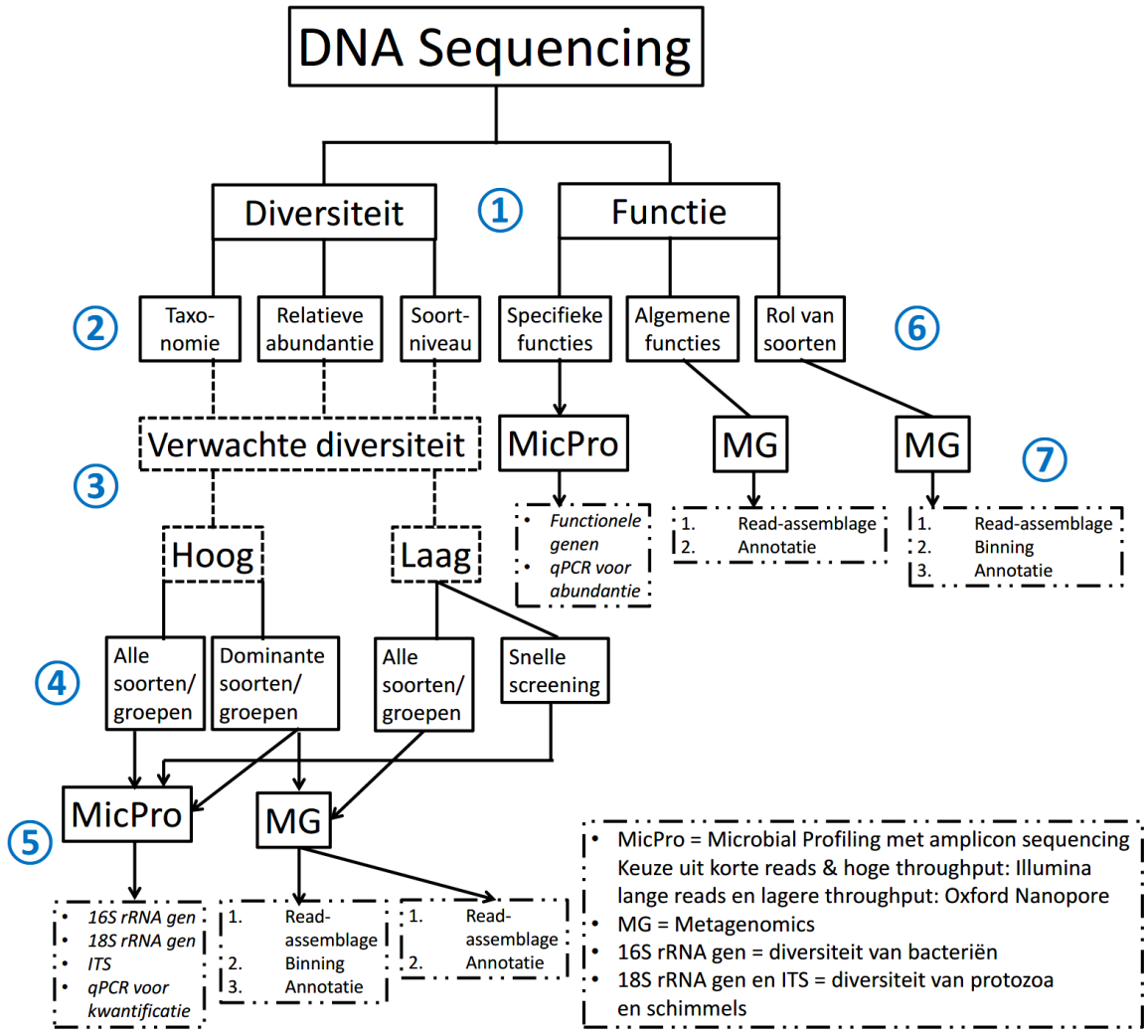
Bij onderzoek naar de implementatiemogelijkheden van metagenomics-analyses is het daarnaast cruciaal om de haalbaarheid voor het opzetten van een bioinformatica-pipeline te inventariseren. Voor het opzetten van een pipeline zoals beschreven in Figuur 17 zijn de volgende factoren van belang:

- Rekencapaciteit van de server: is deze voldoende voor het draaien van de software-tools?
- Bioinformatica-kennis: is aanvullende kennis en training noodzakelijk voor het kunnen uitvoeren van de voorgestelde bioinformatica-pipeline?
- Microbiele fysiologie-kennis: is aanvullende kennis en training noodzakelijk voor het correct uitvoeren en interpreteren van de bioinformatica-output?
- Kracht van de pipeline: levert de voorgestelde pipeline voldoende relevante informatie over processen in drinkwaterzuivering en –distributie?
- Hypothese vs methode: Past de hypothese van het onderzoek bij de gekozen methode? Zie hiervoor ook de flowchart aan het eind van dit hoofdstuk.

Expertise over metagenoom-data-analyse, interpretatie en basiskennis over de verschillende bioinformatica-tools en microbiële fysiologie is binnen KWR aanwezig. Echter, het opzetten en valideren van een dergelijke metagenoom analyse-pipeline neemt veel tijd in beslag. Veel universiteiten beschikken tegenwoordig over een uitgebreide bioinformatica-pipeline en hebben binnenshuis tevens bioinformatici met ervaring voor het uitvoeren van metagenoom-analyses. Een afweging is daarom om voor de verwerking van metagenoom-data de samenwerking met universiteiten te zoeken.
Het voorstel is om de haalbaarheid van metagenoom-analyses volgens het hierboven gedefinieerde schema in een vervolgexperiment vast te stellen. Daarnaast wordt contact gelegd met de afdeling microbiologie aan de Wageningen Universiteit (WUR) en de afdeling Ecologische Microbiologie aan de Radboud Universiteit (RU) waar vanuit KWR contacten zijn.

2.5 Flowchart voor de selectie van DNA sequencing-methoden: NGS versus metagenomics

Om de keuze voor de geschikte DNA sequencing-methode te vereenvoudigen is hieronder een flowchart weergegeven voor DNA-sequencing met hierin de verschillende opties en afwegingen (Figuur 18). Onder de flowchart is een uitgebreide uitleg van de verschillende stappen gegeven.



Figuur 18. Flowchart voor de methoden-keuze tijdens sequencing-projecten. De stappen en keuzes worden hieronder verder toegelicht. Voor verduidelijking van de legenda is in deze figuur "microbial profiling" afgekort als MicPro.

Hieronder worden de keuzes per stap in meer detail toegelicht. Een belangrijke afweging die van tevoren gemaakt dient te worden is of er wordt gekozen voor korte reads met een hoge throughput of juist lange reads met een lagere throughput, maar met een hogere resolutie.

① Allereerst dient een keuze gemaakt te worden of het doel van het sequencing-project is om de diversiteit van de microbiële populatie (wie zit er in het monster?) of de functionaliteit van de microbiële populatie (wat kunnen de aanwezige micro-organismen?) in kaart te brengen.

② Wanneer het doel is om de diversiteit van de microbiële populatie in kaart te brengen kan gekozen worden voor een taxonomische analyse, het berekenen van de relatieve abundantie of identificatie tot op soortniveau, veelal op basis van het 16S rRNA gen.

-Taxonomie: het doel van het experiment is het in kaart brengen van de microbiële diversiteit in een monster. Dit kan grafisch worden uitgewerkt in bijvoorbeeld diversiteitsindexes in boxplots en fylogenetische stambomen.

-Relatieve abundantie: het doel van het experiment is om de relatieve percentages van verschillende groepen micro-organismen die in een monster zitten in kaart te brengen. Dit kan weergegeven worden in bijvoorbeeld bar charts of heat maps.

-Soortniveau: het doel van het experiment is om de aanwezige micro-organismen tot op soortniveau te identificeren. Dit is bijvoorbeeld van belang voor het in kaart brengen van aanwezige opportunistische pathogenen of indicatorsoorten. In nog meer detail kan zelfs de wens bestaan om verschillende microbiële stammen in kaart te brengen.

③ Wanneer de verwachte diversiteit in een systeem erg hoog is (duizenden soorten verwacht) is de resolutie van de gekozen sequencing-methode erg belangrijk. Wanneer de resolutie te laag is, zal een zeer beperkt beeld van het geanalyseerde monster in kaart worden gebracht. Wanneer een monster met een relatief lage verwachte diversiteit (tot enkele honderden soorten) wordt verwacht, kan voor minder diepe sequencing gekozen worden. Wanneer de diversiteit van een systeem nog niet bekend is, valt MicPro, in dit geval een algemene NGS-screening van het 16S rRNA gen als eerste onderzoek aan te raden.

④ Vervolgens dient gekozen te worden of in het onderzoek een volledig beeld van de microbiële diversiteit of bijvoorbeeld alleen van de dominante organismen dient te worden geschetst.

-Hoge diversiteit: in het geval van een hoge diversiteit moet een methode met voldoende throughput worden gekozen. Wanneer het doel bestaat om de volledige diversiteit in kaart te brengen is hierbij MicPro de enige praktisch haalbare methode. Wanneer alleen de dominante micro-organismen in kaart dienen te worden gebracht, kan zowel gekozen worden voor MicPro als metagenomics. Metagenomics heeft minder last van een PCR bias, hetgeen bij NGS studies een effect heeft op de data. Vanwege de hogere kosten is metagenomics voornamelijk waardevol wanneer ook aanvullende vragen over de microbiële functie bestaan.

-Lage diversiteit: voor een snelle en brede screening vormt MicPro de beste en voordeligste optie. Wanneer de vraag bestaat om alle soorten en groepen in een monster met lage diversiteit in kaart te brengen, vormt metagenomics een interessante optie. Ook voor systemen met een lage complexiteit blijft voldoende sequencing met metagenomics cruciaal. De verdere afwegingen hierbij zijn hetzelfde als voor systemen met een hoge diversiteit.

⑤ De keuze voor MicPro-/amplicon-sequencing of metagenomics bepaalt welke antwoorden met de functionele data verkregen kunnen worden.

-MicPro: Bij de MicPro- ofwel de amplicon-methoden wordt gekozen voor de sequencing van het 16S rRNA gen voor het in kaart brengen van de prokaryote diversiteit (bacteriën en archaea) en sequencing van het 18S rRNA gen of de ITS-sequences voor het in kaart brengen van de eukaryote diversiteit (onder andere protozoa en schimmels). Amplicon-data uit “microbial profiling” studies is relatief snel te verwerken en geeft een globaal beeld van de microbiële diversiteit. De methode is geschikt voor het in kaart brengen van alle soorten en groepen micro-organismen en daarbij uiteraard ook voor de dominanten soorten en groepen. De belangrijkste limitatie van de methode is de primer bias, waardoor diversiteit van organismen waarop de primers slecht binden niet goed wordt meegenomen en waardoor tevens de relatieve abundantie niet altijd de werkelijke situatie weergeeft.

-Metagenomics: Bij metagenomics-methoden voor diversiteits-analyses wordt gekozen voor read-assemblage en directe annotatie om een algemeen beeld van de microbiële diversiteit te krijgen. Wanneer de wens bestaat om identificatie tot op soortniveau te doen is het waardevoller om na read-assemblage metagenoom binning en vervolgens annotatie toe te passen.

⑥ Wanneer het doel is om de functies van de microbiële populatie te onderzoeken kan gekozen worden voor het in kaart brengen van specifieke functies van micro-organismen, de algemene functies van een systeem of de specifieke rol en functies van individuele micro-organismen.

-Specifieke functies: wanneer het doel is om specifieke functies van een monster in kaart te brengen, kan een PCR uitgevoerd worden die één enkel of enkele target-genen in een monster amplificeert. Vervolgens kan amplicon sequencing worden uitgevoerd volgens het protocol dat identiek is aan 16S, 18S en ITS sequencing.

-Algemene functies: wanneer het doel is om de algemene functies van een monster in kaart te brengen dient metagenomics te worden toegepast. Hiermee wordt de volledige gen-diversiteit van een systeem in kaart gebracht. Uitdagingen rond de verwachte diversiteit hebben vooral een effect op de volledigheid van de data. Bij een zeer complex systeem is diepere sequencing (soms meerdere Illumina MiSeq runs) nodig om de volledige metabole potentie in kaart te brengen.

-Rol van soorten: wanneer het doel is om de metabole potentie (mogelijke rol) van specifieke microbiële soorten in kaart te brengen, dient metagenomics te worden toegepast. Hierbij spelen dezelfde uitdagingen rondom de sequencing-diepte als hierboven zijn genoemd.

⑦ De keuze voor NGS-sequencing van specifieke genen of van het metagenoom bepaalt welke antwoorden met de functionele data verkregen kunnen worden.

-Specifieke genen: Sequencing van specifieke functionele genen wordt uitgevoerd voor specifieke microbiële processen (bijvoorbeeld het *amoA* gen dat codeert voor een deel van het ammonium monooxygenase-eiwit dat ammonium oxideert). De methode voor sequencing van de functionele genen is vergelijkbaar als bij sequencing van het 16S en 18S rRNA gen of de ITS-sequenties. Wanneer gekozen wordt voor Illumina sequencing wordt een kort fragment van het gen van vaak enkele honderden basenparen geselecteerd. Ook hiermee kan vervolgens naar de diversiteit van bijvoorbeeld ammonium-oxiderende micro-organismen worden gekeken.

-Metagenomics: Bij metagenomics-methoden wordt de gehele genoomdiversiteit van een monster gesequenced. De resolutie van de data hangt af van de complexiteit van het monster en de gekozen diepte van de sequencing.

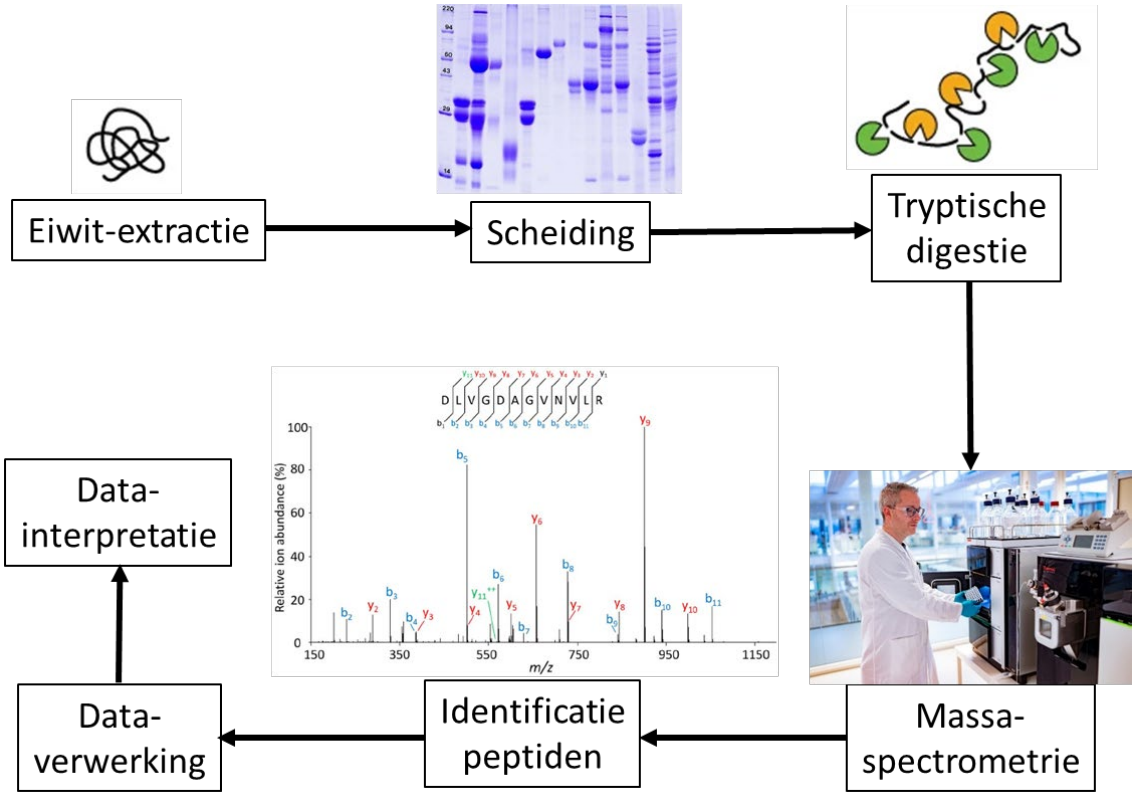
3 Het monitoren van biochemische processen met metaproteomics

3.1 De principes van metaproteomics

Bij metaproteomics worden alle aanwezige eiwitten in een monster geanalyseerd. De eiwitten zijn het product van de tot expressie gebrachte genen in een organisme (Figuur 1). Onder het metaproteoom vallen onder andere enzymen, ofwel de eiwitten die specifieke chemische reacties katalyseren. Metaproteomics geeft daarmee een beeld van de activiteiten en functies die op dat moment worden uitgevoerd. Bij de analyse van microbiële populaties gaat dit een stap verder dan metagenomics, waarbij alleen het metabole potentieel in kaart kan worden gebracht. De metagenomics-data vormt in de regel wel de basis voor de verwerking en interpretatie van metaproteomics-data.
De potentie van metaproteomics is eerder kort besproken in de nieuwsbrief kansen van nieuwe moleculair-microbiologische methoden in de drinkwatersector van juni 2019 (paragraaf 3.2 van deze nieuwsbrief).

Binnen dit deel van de bureaustudie wordt een verkenning gedaan voor de toepassing van metaproteomics in de drinkwaterzuivering en -distributie. De centrale vraag is of met metaproteomics voldoende waardevolle informatie over de microbiologische processen kan worden verkregen.
Het proces van metaproteomics is nog relatief onbekend en niet eerder uitgewerkt in eerdere onderzoeken bij KWR. Daarom wordt allereerst in 3.2 (3.2.1 tot en met 3.2.6) ingegaan op de verschillende stappen in metaproteomics-analyses (Figuur 19). Hierbij wordt voor elke stap kort ingegaan op de methodieken, afwegingen en limitaties voor toepassing in de drinkwatersector. Vervolgens worden in 3.3 de uitdagingen en de meerwaarde van metaproteomics voor toepassing in de drinkwaterproductie geëvalueerd. In 3.3.1 worden vervolgens de implementatiemogelijkheden voor metaproteomics in de drinkwaterzuivering en -distributie uitgewerkt en toegelicht.

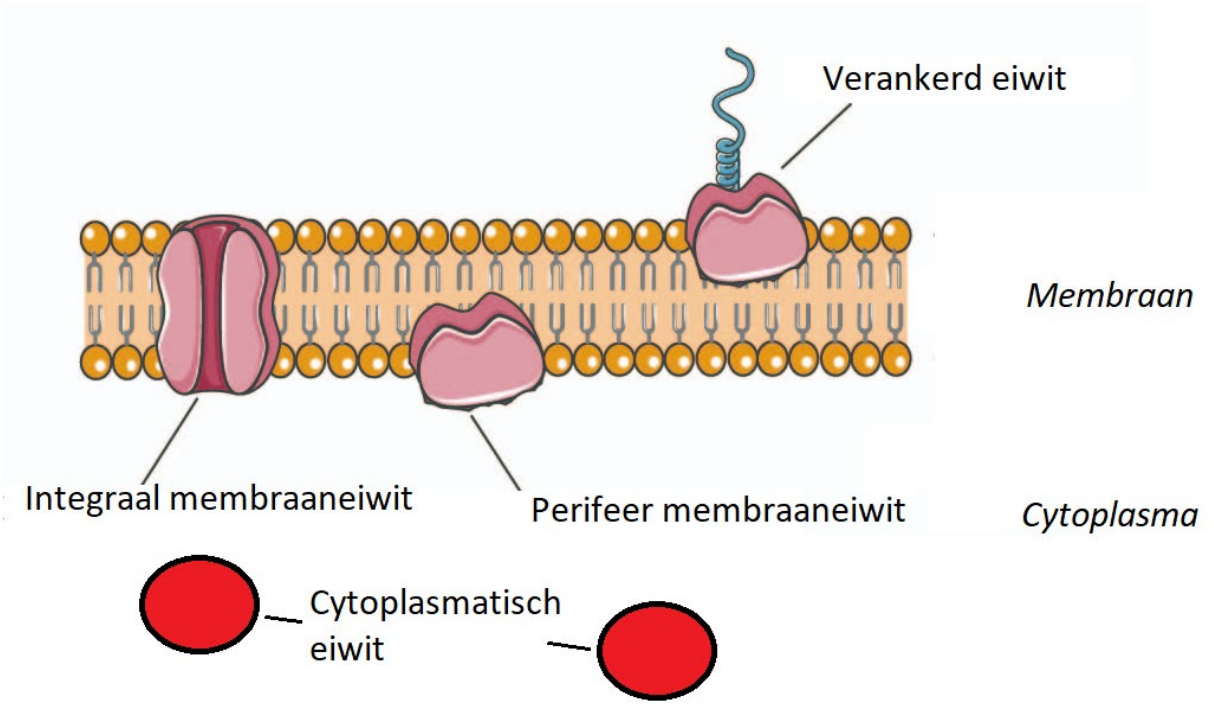
3.2 De stappen in een metaproteoom-analyse



Figuur 19. Overzicht van de stappen in metaproteoom-analyses.
Overzicht van de verschillende stappen in een metaproteoom-analyse van eiwit-isolatie tot data-interpretatie. Op deze verschillende stappen wordt in de paragrafen 3.2.1 tot en met 3.2.7 ingegaan.

3.2.1 Eiwit-extractie

De eerste stap van de metaproteomics workflow is de extractie van eiwitten uit een monster (Figuur 19). Hiervoor is het allereerst van belang om inzicht te hebben waar in de cel de eiwitten zich bevinden. Op basis van de lokalisatie worden eiwitten verdeeld in de categorie I) membraangebonden eiwitten en II) cytosolische of cytoplasmatische eiwitten (Figuur 20). Membraangebonden eiwitten zijn verankerd of geïntegreerd in het celmembraan. Cytoplasmatische eiwitten bevinden zich vrij in het cytoplasma, ofwel de celvloeistof. Dit heeft consequenties voor de eiwit-extractie.



Figuur 20. Overzicht van membraangebonden eiwitten en cytoplasmatische/cytosolische eiwitten.

Bij metaproteomics is het doel om een zo breed mogelijk beeld van de eiwitsamenstelling van een monster te krijgen. Het is daarom gewenst om zowel de membraangebonden als de cytoplasmatische eiwitten te extraheren. De membraangebonden eiwitten vormen een complex met membraanlipiden en overige membraanstructuren. Hierdoor zijn speciale chemische en/of mechanische behandelingen van een monster nodig om deze eiwitten ook in de analyse mee te kunnen nemen. Voor cytoplasmatische eiwitten is dit niet nodig, aangezien deze eiwitten al in oplossing in een cel voorkomen. Voor deze eiwitten is het van belang dat deze selectief uit de oplossing kunnen worden gehaald. Voor metaproteoom-analyses is het van belang om protocollen te selecteren waarbij zo veel mogelijk eiwitten en daarmee zowel een semi-kwantitatief beeld als zo breed mogelijke diversiteit aan eiwitten uit een monster kan worden gehaald.
Belangrijk om te realiseren is dat de extractie-efficiëntie van eiwitten laag is (<5%), waardoor de kans op een extractiebias hoog is. Hier wordt kort op ingegaan onder het kopje 'Uitdagingen bij eiwit-extracties' onderaan deze paragraaf.
Er is een groot aantal verschillende protocollen beschikbaar om eiwitextracties uit te voeren (zie voor een overzicht van verschillende kits bijvoorbeeld: Protein Extraction Kits | Biocompare). Qua methodiek zijn deze in te delen in I) mechanische methoden II) chemische methoden en III) op **affiniteit** gebaseerde methoden. Vaak worden tijdens eiwit-extracties combinaties van deze drie methoden toegepast. Op de principes van de drie verschillende methoden wordt hieronder ingegaan.

I) Mechanische methoden

Voor de extractie van eiwitten uit biologisch materiaal worden vaak mechanische methoden als **sonicatie**, herhaalde vries-droogcycli, **homogenisatie** met hoge druk (de zogeheten **French press**-methode) of **bead-beating** met kleine glasparels gebruikt. Deze methoden zijn er op gericht om cellen open te breken en de in cellen aanwezige eiwitten in oplossing te krijgen.
Een specifiekere mechanische methode die kan worden om verschillende eiwitfracties te isoleren is **differentiële centrifugatie**. Deze methode maakt gebruik van verschillende centrifugatiestappen, snelheden en tijden om de eiwitten in een specifieke fractie opgezuiverd te krijgen. In combinatie met centrifugatie kunnen chemicaliën worden gebruikt om de eiwitten te laten **precipiteren**. Vaak worden hier in de praktijk hoge zoutconcentraties of hoge/lage pH-waarden voor gebruikt. Het bijeffect van deze methoden is dat de eiwitten hun natieve staat (de vorm waarin ze van nature in cellen voorkomen) verliezen. Voor metaproteomics-toepassingen is dit echter niet

noodzakelijk.
Ook kan door toepassing van chemicaliën, zogeheten **detergentia**, worden gezorgd dat membraangebonden eiwitten in oplossing komen. Vooral voor metaproteomics is het toevoegen van een dergelijke stap belangrijk om ook de membraangebonden fractie van het metaproteoom mee te nemen.

II) Chemische methoden

Er bestaan verschillende chemische methoden voor de extractie van eiwitten. Veelgebruikte chemicaliën zijn bijvoorbeeld Triton X-100, Trizol, guanadine hydrochloride (GuHCl) en citraat-fenol) of enzymatische methoden worden gebruikt.
Onder chemische methoden valt tevens de toepassing van enzymen. Enzymen die worden toegepast voor eiwit-extractie zijn lysozym, cellulase en protease (proteinase). Lysozym is een enzym dat het peptidoglycaan, een structurele component in bacteriële celwanden afbreekt. Hierdoor verliezen bacteriën hun integriteit en breken ze open (lysis). Cellulase is een enzym dat cellulose afbreekt. Cellulose komt naast planten ook voor in de celwand van enkele bacteriesoorten zoals *Rhizobium* en *Pseudomonas*. Afbraak leidt ook hier tot lysis.
Protease katalyseert de afbraak van eiwitten. Door de afbraak van eiwitten in de celwand en celmembraan lijdt ook een proteasebehandeling tot lysis.
Zoals hierboven genoemd, wordt vaak een combinatie van chemische en mechanische methoden gebruikt. Hierbij dient de chemische methode specifiek om membraangebonden eiwitten los te krijgen.

III) Op affiniteit gebaseerde methoden

De lading van de verschillende aminozuren kunnen gebruikt worden voor de selectieve opzuiveringen van eiwitten. Veelgebruikte technieken maken gebruik van hydrofobe kolommen en ionuitwisselingskolommen. Met ionuitwisselingschromatografie kunnen eiwitten op basis van lading gescheiden worden.
In het algemeen worden op affiniteit gebaseerde methoden gebruikt voor de opzuivering van specifieke eiwitten en eiwitgroepen. De methode is bewerkelijk en kostbaar. Daarmee is de techniek niet direct relevant voor de toepassing bij metaproteomics-studies.

Uitdagingen bij eiwitextracties

De belangrijkste limitatie bij eiwitextracties in het algemeen, en zeker voor drinkwatermonsters, is de totale hoeveelheid beschikbaar eiwitmateriaal. In de praktijk geldt namelijk dat een hogere concentratie eiwitmateriaal in de meeste gevallen een betere resolutie van de resultaten oplevert.
Zoals kort aangestipt bij de behandeling van extractiemethoden, resulteert de in het algemeen lage extractie-efficiëntie (<5%) van de verschillende methoden in een potentieel grote bias. Uit een vergelijkingsstudie van verschillende eiwit-extractiemethoden blijkt verder dat er een zeer beperkte overlap is in de resulterende metaproteoom-data (1,9% overlap in eiwitsequenties, 8,3% overlap in eiwitfamilies) (Leary et al. 2013). De gekozen extractiemethode heeft dus een grote invloed op het resultaat. Het is daarom van belang om voor de vergelijkbaarheid tussen verschillende analyses een goede afweging voor de meest geschikte extractiemethode te maken en vervolgens eenzelfde extractiemethode aan te houden.
Binnen drinkwater-analyses vormen de drinkwatermonsters door de lage concentratie biologisch materiaal een mogelijke beperking. Uit biofilmmonsters en sediment kan daarentegen relatief eenvoudig voldoende eiwit-materiaal gehaald worden. Voor drinkwater geldt dat voor een goede proteoom-analyse minimaal 1 liter dient te worden gefilterd. Er geldt echter, hoe meer eiwit-materiaal, hoe beter de resolutie van de uiteindelijke analyse (Martin Pabst (Microbial Proteomics Group, TU Delft), persoonlijke communicatie). Een realistische aanname is daarom dat minimaal enkele liters gefilterd dienen te worden voor een gedegen metaproteoom-studie. Dit is in de praktijk goed haalbaar. Voor de ontwikkeling van een protocol voor het verzamelen van eiwitmateriaal zal van tevoren contact gezocht worden met de Microbial Proteomics Group aan de TU in Delft.
Contaminatie speelt bij eiwitwerk tevens een grote rol. Vooral tijdens de extractie kan besmetting van het monster met menselijke eiwitten, vooral afkomstig van de huid, voorkomen. Alhoewel dergelijke besmettingen tijdens de data-analyse relatief eenvoudig te verwijderen zijn, heeft het een negatieve invloed op de resolutie van de analyse.

Het is daarom van belang dat de eiwit-extractie maar ook de vervolgstappen zo schoon mogelijk worden uitgevoerd.
Aanvullend geldt dat eiwitten stabiel zijn voor de duur van enkele uren. Het is daarom van belang dat de te analyseren monsters zo snel mogelijk worden behandeld, zodat er geen modificaties in het metaproteoom optreden. De grootste uitdaging ligt hier bij de verwerking van drinkwatermonsters, al is dit binnen enkele uren goed praktisch uitvoerbaar.

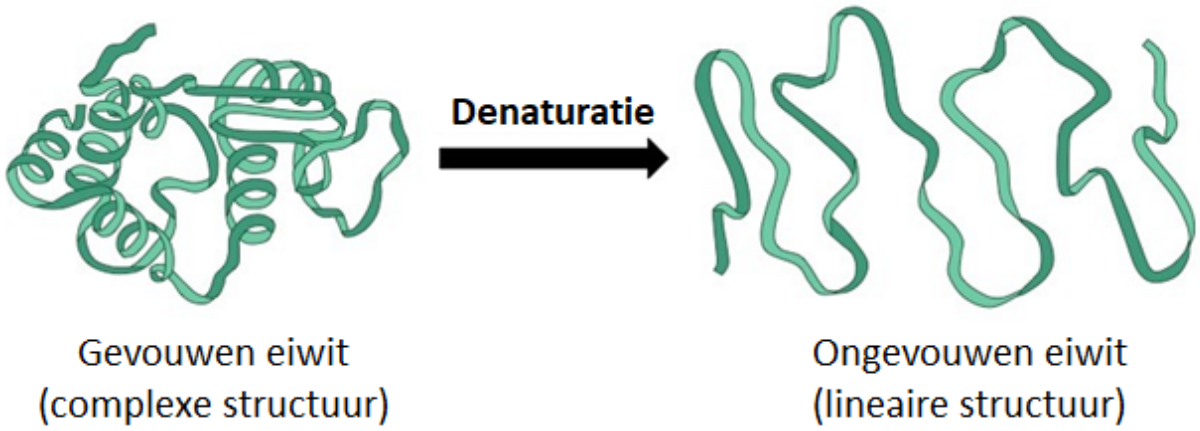
Uit overleg met Martin Pabst (Microbial Proteomics Group, TU Delft) volgt dat de optimalisatie van eiwitextractiemethoden voor drinkwater een langdurig en ingewikkeld proces is. Voor implementatie op korte termijn lijkt het daarmee aan te raden om de volledige metaproteomics workflow vanaf het moment van eiwitextractie te laten uitvoeren door een laboratorium met ervaring op het gebied van metaproteomics en specifieke ervaring op drinkwatermonsters.

3.2.2 Eiwitscheiding

Een eiwitscheidingsstap is noodzakelijk om de complexiteit in het monster te verlagen voor de uiteindelijke massaspectrometrische analyse (MS-analyse) (3.2.4). De MS-analyse dient namelijk te worden uitgevoerd op kleinere eiwitfracties om voldoende resolutie te behalen. De eiwitscheidingsstap wordt standaard op eiwitgrootte uitgevoerd met SDS-PAGE gel-electroforese.

SDS-PAGE voor eiwitscheiding op grootte

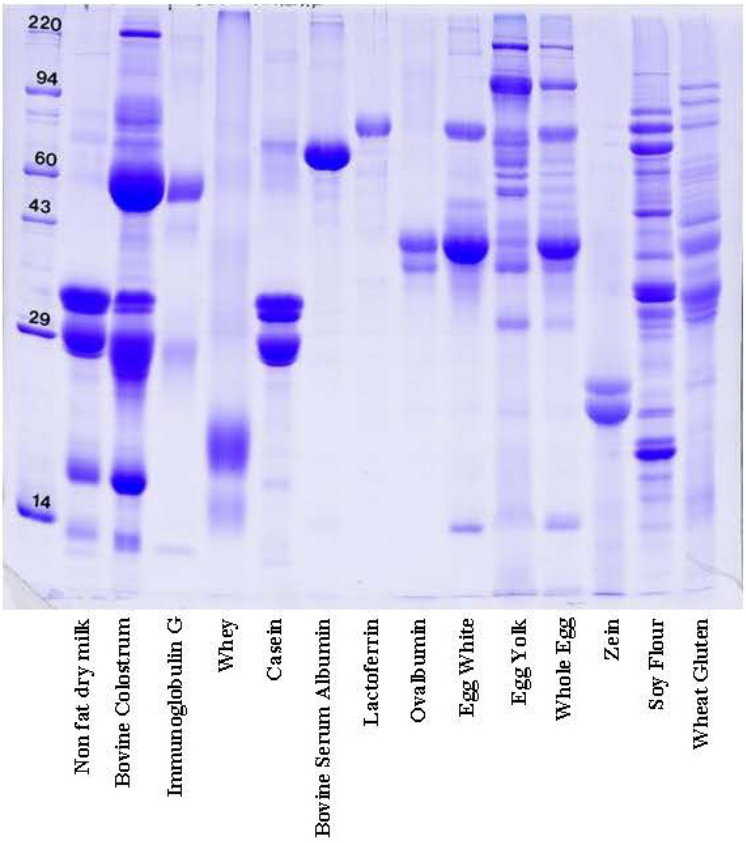
SDS-PAGE gel-electroforese is de meest breed toegepaste methode. Bij deze methode wordt in de eerste stap het eiwitmengsel gedenatureerd in natriumdodecylsulfaat (SDS, sodium dodecyl sulfate) bij 70 °C. Bij denaturatie worden de verschillende zwakke interne bindingen van een eiwit verbroken. Hierdoor verliezen de eiwitten hun complexe structuur (ook wel de natieve conformatie genoemd) en komen ze in een lossere, meer lineaire structuur (Figuur 21). Deze structuur is nodig om de scheiding op basis van grootte te kunnen doen.



Figuur 21. Het proces van denaturatie voor het verkrijgen van eiwitten in de ongevouwen, lineaire structuur.

Vervolgens wordt het gedenatureerde eiwitmengsel gescheiden door middel van polyacrylamide gel-electroforese (PAGE). Vervolgens wordt een kleurstof (meestal coomassie blue) gebruikt om de eiwitten op de gel zichtbaar te maken (Figuur 22). Op basis van de aangekleurde gel kan een eerste beeld worden verkregen van de grootteverdeling van het eiwitmonster. In het voorbeeld van Figuur 22 is meest links in de gel een aantal banden te zien van een ladder, ofwel een standaard eiwitmengsel met daarin eiwitten van bekende grootte. De grootte van deze fragmenten is weergegeven in kilodalton ofwel kDa (hierbij geldt 1 kDa = 1 kg/mol). Ook zijn er vaak enkele

zeer intens gekleurde banden in de gel waarneembaar die een indicatie geven dat bepaalde eiwitten in het monster zeer abundant zijn.



Figuur 22. Voorbeelden van eiwitscheiding en aankleuring met coomassie blue met de SDS-PAGE methode.
In elke zogeheten laan is het eiwitprofiel van verschillende monsters, zoals onderaan beschreven, te zien. Geheel links is een eiwitladder toegevoegd die de referentiemaat in kilodalton (kDa) weergeeft. In het algemeen geldt dat kleinere eiwitten sneller door de gel migreren en dus onderaan in de gel terechtkomen. Bron: Kendrick Laboratories.

Van de zogeheten SDS-PAGE gel kunnen vervolgens secties gemaakt worden voor verdere analyse. Meestal worden hiervoor circa 20 secties van een enkel monster gemaakt. Het maken van meer secties is vaak praktisch niet haalbaar (de gelfracties worden dan erg klein) en tevens wordt de concentratie eiwitten in de fracties te laag.

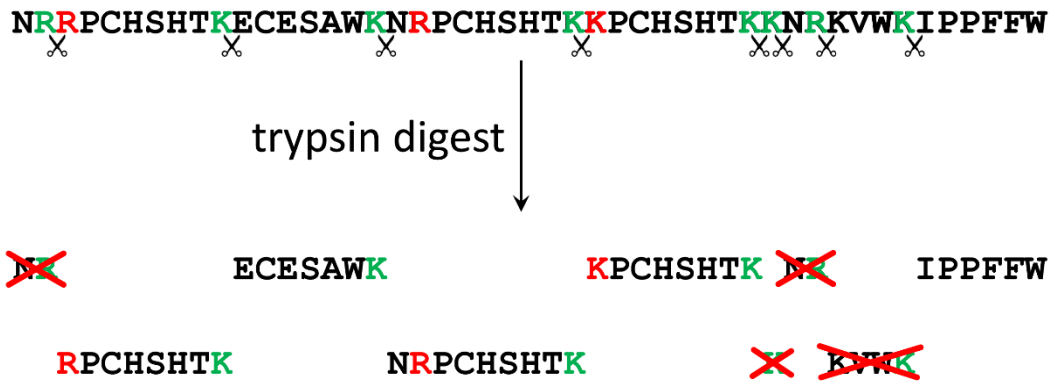
Uitdagingen bij eiwitscheiding

In het algemeen geldt dat door het toepassen van een SDS-PAGE gel er in verhouding relatief grote hoeveelheden eiwit nodig zijn (5-20 µg per laan). Een gedeelte van het eiwitmateriaal zal in het SDS-PAGE proces verloren gaan. Daarnaast zal de gel voor de vervolgstappen worden opgedeeld in fracties, waardoor het noodzakelijk is dat elke fractie voldoende eiwitmateriaal bevat. Terugkomend op voor drinkwater specifieke monsters zal dit voor de analyse van biofilms en sedimenten geen probleem opleveren. Voor drinkwater is het hierdoor van belang om minimaal enkele liters als inputmateriaal te gebruiken.
Limitaties die specifiek optreden bij de SDS-PAGE gel-electroforese worden vooral gevormd door eiwitten en **eiwit-complexen** die niet in staat zijn om de SDS-PAGE gel in te migreren. Op de gel is deze fractie aan het eind van de analyse zichtbaar als een dikke, blauwe band bij het laadslot van het specifieke monster. Naast deze eiwitcomplexen blijven in deze band eventuele onzuiverheden (mogelijke celresten) achter. Daarmee is de SDS-PAGE gel dus tevens een belangrijke zuiveringsstap waarin niet-eiwitfracties verwijderd worden.
Het is van belang om goed in de gaten te houden of er een dikke gekleurde band bij het laadslot optreedt, aangezien dit een grote bias in de vervolganalyse kan opleveren. Indien dit het geval is, zijn aanpassingen in de

voorbereidingen van het monster vóór de SDS-PAGE run noodzakelijk.

3.2.3 Tryptische digestie

Na de eiwitscheiding wordt op de eiwitfracties een digestie toegepast. Bij **tryptische digestie** wordt het enzym trypsine, een eiwit-hydrolyserend enzym (een protease), gebruikt om de eiwitten uit het monster in kleine stukken, ofwel peptiden, te knippen (Figuur 18). Deze peptiden hebben een lengte van minimaal 2 tot maximaal 50 aminozuren. Tryptische digestie is noodzakelijk omdat volledige eiwitten te groot zijn om direct met massaspectrometrie geanalyseerd te kunnen worden.



Figuur 23. Overzicht van het werkingsmechanisme van tripsine.
Trypsine is een enzym dat meestal aan het einde van de aminozuren lysine (K) en arginine (R) knipt, met uitzondering van de situatie waarin beiden worden ogevolgd door proline (P). Figuur afkomstig van Unipept (Unversiteit Gent). Peptiden groter dan 50 aminozuren of kleiner dan 5 aminozuren vallen in de regel buiten de detectielimiet van de meetapparatuur gebruikt voor massaspectrometrie (MS) en vallen daardoor af.

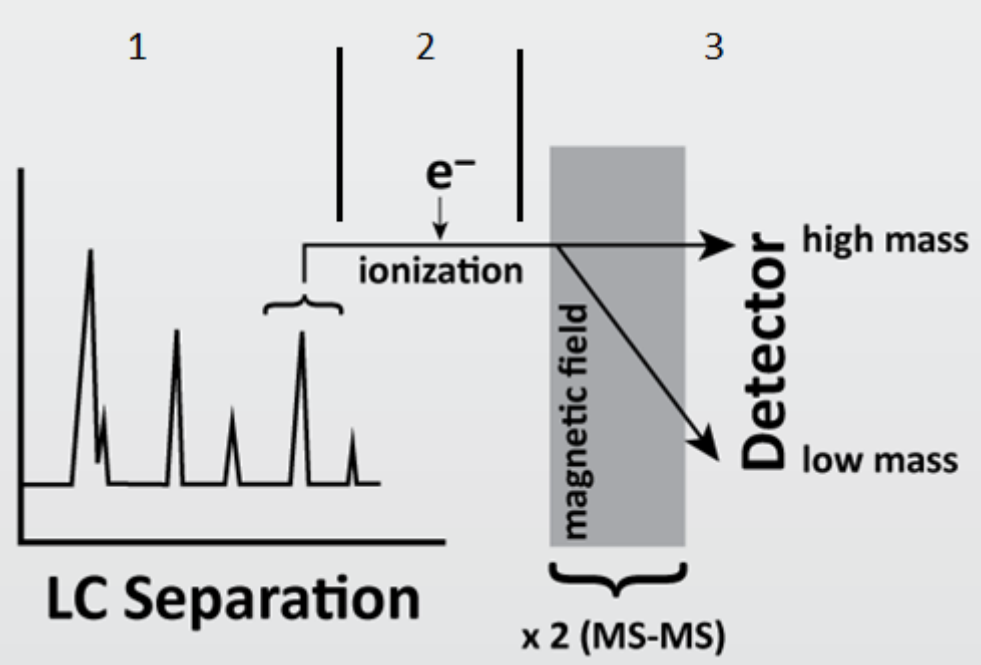
Uitdagingen bij tryptische digestie

De aanwezigheid van interfererende stoffen is de belangrijkste factor die een negatief effect kan hebben op de digestiestap van eiwitten. Dit komt doordat de digestiestap een enzymatische reactie is. De enzymactiviteit wordt sterk beïnvloed door de aanwezige omstandigheden (o.a. pH, temperatuur, concentratie zouten, chemicaliën etc.). De eiwitscheiding op een SDS-PAGE-gel biedt hierbij overigens een belangrijke zuiveringsstap waarbij een groot gedeelte van mogelijk interfererende stoffen wordt weggezuiverd.
Daarnaast kunnen sterk gevouwen eiwitten en sterke eiwit-interacties problemen opleveren met de digestie. In principe worden deze structuren en bindingen verwijderd tijdens het SDS-PAGE-proces. Toch is, zeker in een complex eiwit-mengsel, de kans groot op een kleine fractie eiwitten die hierdoor niet voldoende denatureert. Het gevolg is dat digestie-eiwitten deze fractie minder goed kunnen opknippen, doordat de structuren minder toegankelijk zijn. Daarmee bestaat er een bias voor deze groep van eiwitten.

3.2.4 Massaspectrometrie (MS)

Na de tryptische digestie wordt op het peptidenmengsel massaspectrometrie (MS) uitgevoerd om de peptiden te identificeren. De MS-analyse bestaat uit een drietal stappen (Figuur 24):
1. Scheiding van het peptidenmengsel met vloeistofchromatografie / **liquid chromatography** (LC).
2. Ionisatie van de peptidenmengsel-fracties (“ionization”).
3. Analyse van de peptidenmengsel-fracties met massaspectrometrie / **mass spectrometry** (MS).

Op de drie stappen wordt hieronder in meer detail ingegaan.



Figuur 24. Overzicht van de LC-MS-methode.
De LC-MS-methode bestaat uit 3 stappen: 1. Scheiding met liquid chromatograpy (LC); 2. Ionisatie van de peptidenmengsel-fracties; en 3. Massaspectrometrie (MS), met in dit voorbeeld een dubbele MS-stap (MS-MS).

Stap 1: Liquid chromatography (LC) voor scheiding van het peptidenmengsel

Het door tryptische digestie verkregen peptidenmengsel heeft een te hoge complexiteit voor een directe analyse met MS. Om de complexiteit te verlagen en voldoende resolutie te bereiken, is de eerste stap in de massaspectrometrische analyse daarom een scheiding van de eiwit-oplossing met liquid chromatography (LC) (Figuur 24). In de meeste gevallen wordt bij LC-MS gebruik gemaakt van de **reverse phase** (**RP**) modus voor chromatografie. Hierbij wordt gebruik gemaakt van een **apolaire** (hydrofobe, waterafstotende) vaste fase en een **polaire** (hydrofiele, wateroplosbare) vloeistoffase. De scheiding van het peptidenmengsel vindt bij de RP modus plaats op basis van de verschillen in affiniteit voor de apolaire en polaire fase. Onder hoge druk worden verschillende **sorptie-desorptie**stappen (ofwel stappen van afwisselend binden en loslaten) uitgevoerd.
Tijdens deze LC-stap worden de peptiden gescheiden op basis van hun chemische eigenschappen, waardoor dezelfde en soortgelijke peptiden geclusterd worden. Dit heeft een positief effect op de resolutie van de MS-analyse.
De chromatograaf is gekoppeld aan de **electrospray-ionisatie** (**ESI**) unit waar stap 2, de ionisatie van de peptiden plaatsvindt.

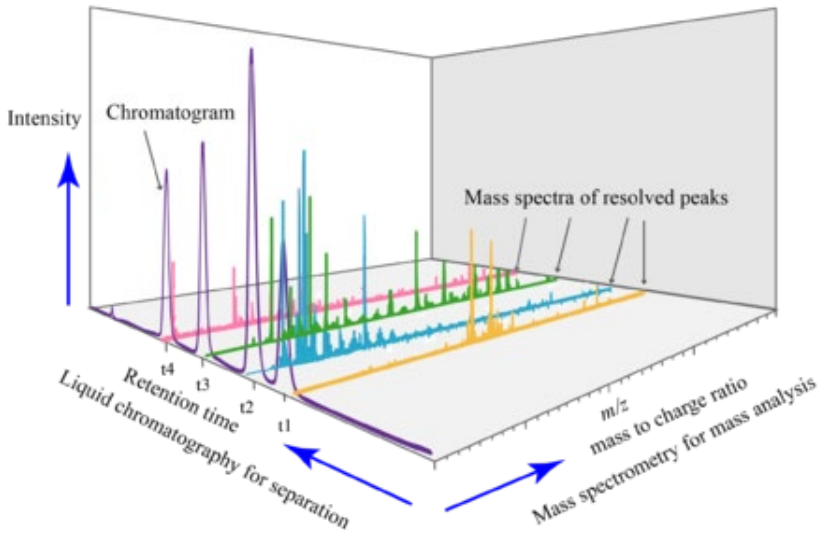
Stap 2: ionisatie

Met ionisatie worden de peptiden uit oplossing omgezet in geladen biomoleculen in de gasfase. Meestal wordt hiervoor “electrospray ionization” (ESI) toegepast. Hierbij wordt een hoog voltage toegepast op de LC-vloeistof om aerosolen te creëren. Vervolgens kunnen deze geladen moleculen in een elektrisch/magnetisch veld worden gekarakteriseerd door de massaspectrometer.

Stap 3: Massaspectrometrie (MS)

Met MS wordt de zogeheten “**mass to charge ratio**” weergegeven als “**m/z**” van de in stap 2 gevormde geïoniseerde deeltjes. Het getal dat hiervoor wordt weergegeven is de massa van een peptide-ion gedeeld door de lading van het peptide-ion. Doordat de peptiden in individuele LC-pieken een verschillende massa hebben en tevens een verschillende lading krijgen tijdens de ionisatie, biedt de “mass to charge ratio” een goede onderscheidende dimensie.

De uiteindelijke massaspectrometrische analyse heeft dus twee hoofddimensies: de LC-dimensie en de MS-dimensie. Een voorbeeld van een uiteindelijk resultaat is weergegeven in Figuur 25. In dit figuur is goed te zien hoe de LC-stap door de fractionering van het peptidenmengsel zorgt voor een verhoogde resolutie van de massaspectrogrammen uit de MS-analyse.



Figuur 25. Voorbeeld van een LC-MS spectrum.
Op de as van de 'Retention time' is het resultaat van de eerste scheiding door liquid chromatography (LC) te zien. Op de as van de 'mass to charge ratio' (m/z) zijn de verkregen massaspectrogrammen van de verschillende pieken uit de massaspectrometrie (MS)-analyse zichtbaar.

Variaties in MS-analyses

Tijdens de massaspectrometrische analyse wordt vaak gekozen voor een dubbele MS-analyse, ook wel tandem MS, MS/MS of MS^2 genoemd. Bij deze opzet worden twee massaspectrometers achter elkaar gekoppeld met als doel het detectievermogen en de specificiteit verder te vergroten. Bij tandem MS hangt de tweede MS vaak af van de resultaten van de eerste MS. Er kan aanvullend gekozen worden om de twee MS-apparaten in een andere modus te laten draaien, waardoor meer informatie over de peptiden verkregen kan worden. Er zijn tal van verschillende combinaties mogelijk en binnen gespecialiseerde laboratoria voor metaproteomics bestaan vaak geoptimaliseerde opstellingen voor de analyse.

Uitdagingen bij massaspectrometrie

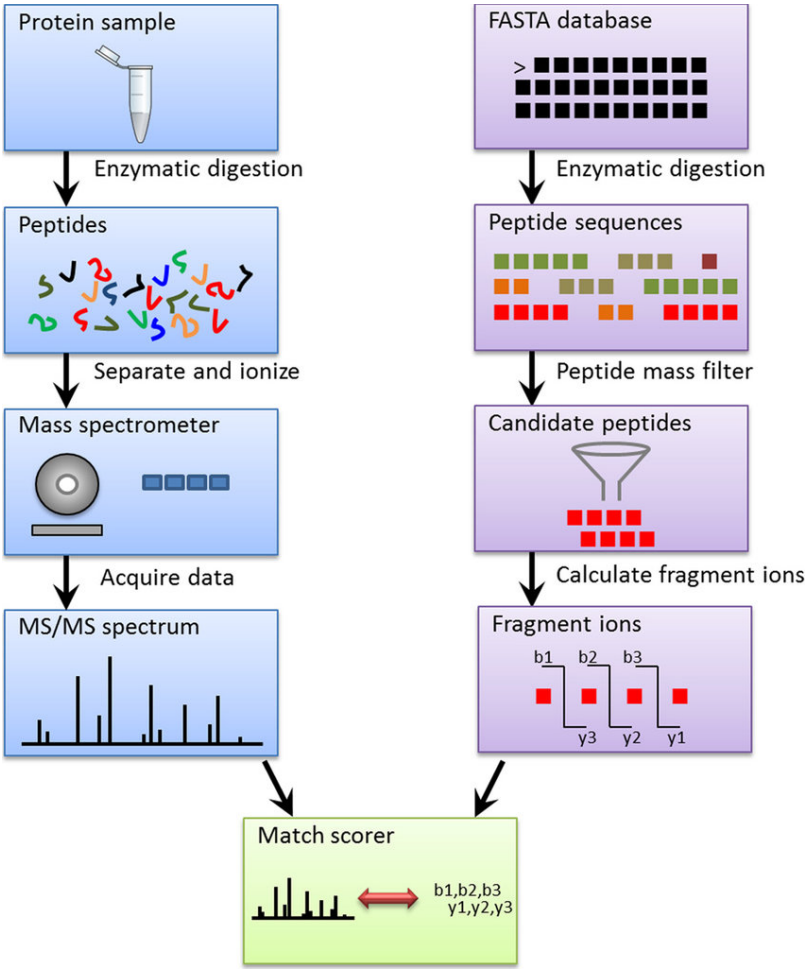
Alhoewel grote ontwikkelingen binnen de massaspectrometrie hebben plaatsgevonden, blijft kwantificatie van peptiden en daarmee eiwitten met deze technologie een grote uitdaging.
Een relatieve kwantificatie van eiwitten in een monster is relatief eenvoudig uit te voeren door de abundantie van specifieke eiwitten in verschillende monsters met elkaar te vergelijken. Dit wordt dan weergegeven als relatieve log-ratio per eiwit. Hierdoor kan een beeld verkregen worden van specifieke monsters met hoge en lage abundanties van bepaalde eiwitten. Dit is de methode die in de praktijk het meest wordt toegepast.
Voor absolute kwantificatie wordt vaak gebruik gemaakt van radioactief of chemisch labelen van eiwitten/peptiden. Dit is echter bewerkelijk en praktisch niet haalbaar voor metaproteomics. Hiervoor zijn ook labelvrije, zogeheten '**peptide-centric**' benaderingen. Hierbij wordt het aantal gedetecteerde peptiden per eiwit geteld of wordt gekeken naar de relatieve abundantie van specifieke peptide-ionen (dit zijn de piek-intensiteiten uit de LC-MS data).
Ook deze kwantificatie is niet biasvrij. Dit komt doordat verschillende peptiden een verschillende ionisatie-efficiëntie hebben en door verschillen in detectiegevoeligheid van verschillende peptiden. Het is van belang om deze limitaties en uitdagingen in het achterhoofd te houden bij de toepassing van kwantificatie op LC-MS data.

3.2.5 Eiwitidentificatie

Uiteindelijk dient de MS-data omgezet te worden in eiwitsequenties. Hiervoor worden allereerst softwaretools gebruikt om de verschillende peptiden te identificeren. Deze tools zijn onderverdeeld in database-gebaseerde tools en *de novo* tools. De database-gebaseerde tools maken gebruik van genoomdata van waaruit alle mogelijk aanwezige aminozuursequenties van de peptiden kunnen worden afgeleid. Hiervoor kan gebruik worden gemaakt van algemene databases, zoals de UniProt database, maar ook van de in het experiment gegenereerde metagenoom-data. Het voordeel van het gebruik van metagenoom-data is dat zeker voor relatief onbekendere milieus meer peptiden en eiwitten kunnen worden geïdentificeerd. In de praktijk geldt dat voor relatief onbekende monsters de database-gebaseerde tools de beste resolutie opleveren. Bij *de novo* peptide sequencing vindt massaspectrometrie plaats zonder enige voorkennis over de identiteit van de verschillende eiwitten. Dit is minder geschikt voor monsters waar weinig over bekend is. Hier wordt daarom niet verder op ingegaan.

Van signaal naar peptide met database-gebaseerde tools

De massaspectrometrische analyse geeft een zeer accurate massa van de verschillende peptiden. Op basis van deze massa kan zogeheten "peptide fingerprinting" worden uitgevoerd. Hiervoor worden de verkregen massa's in een database met fragmenten verkregen uit de digestie en analyse van bekende eiwitten. Wanneer een groot aantal peptiden een hit geeft op een eiwit dat in de database aanwezig is, kan worden vastgesteld dat het betreffende eiwit waarschijnlijk in het monster aanwezig was. Hiervoor is het echter noodzakelijk dat er informatie voor de eiwitten aanwezig is in de database. Dit is echter niet altijd het geval.
Voor veel metaproteoom-werk geeft het gebruik van metagenoom-data als referentie de hoogste resolutie. Wanneer metagenoom-data beschikbaar is, wordt de genoom-data (nucleotiden) omgezet naar aminozuur- en eiwitsequenties. Dit is de input-data die vervolgens *in silico* op de computer dezelfde stappen doorloopt als het echte eiwitmonster, van enzymatische digestie tot detectie van het massaspectrum (Figuur 26).



Figuur 26. Overzicht voor de identificatie van peptiden uit een eiwitmonster (links) en een metagenoom referentiedataset (rechts). Links de workflow van eiwit-extractie naar massaspectrum (MS/MS spectrum in dit voorbeeld) en rechts de workflow waarbij op basis van een metagenoom (hier FASTA database genoemd) een in silico digestie op de computer wordt uitgevoerd, waarna softwarematig kandidaat peptiden worden geselecteerd en wederom in silico fragment-ionen worden gegenereerd. Dit wordt vervolgens langs de massaspectrometrie-data gelegd om de meest passende match te vinden. Figuur afkomstig uit de studie van Eng en co-auteurs (Eng et al. 2011).

Wanneer een relatief groot aantal eiwitten onbekend is, zoals vaak het geval is bij metaproteomics, is het gebruik van een zogeheten peptide spectral library interessant. Dit is een handmatig samengestelde en opgeschoonde database van peptidenspectra. Hiervoor is het echter noodzakelijk dat een goede peptiden-database aanwezig is. Juist voor onbekendere monsters is dit momenteel niet haalbaar. Voor toepassing in drinkwateronderzoek is dit daardoor momenteel nog niet mogelijk. In dit geval wordt “peptide fingerprinting”, zoals hierboven beschreven toegepast.

3.2.6 Data-interpretatie

Het doel is om uiteindelijk een zo groot aantal eiwitten in een monster te kunnen identificeren en waar mogelijk iets over relatieve abundanties te zeggen. Er is een groot aantal softwaretools beschikbaar voor de analyse en interpretatie van de massaspectrometrische datasets. Een aantal tools die veelal worden toegepast voor (meta)proteoom-analyses is weergegeven in Tabel 5.

Tabel 5. Overzicht van een aantal softwaretools voor de verwerking en interpretatie van massaspectrometrische data.

Softwaretool	Kenmerken	Opmerkingen
MaxQuant	Open-source kwantitatieve proteomics-tool voor de analyse van grote massaspectrometrische datasets.	In het programma zit een zoekfunctie en visualisatietool. Aanvullende statistische analyses mogelijk met de Perseus tool.
MetaProteomeAnalyzer	Open-source tool ontwikkeld voor metaproteoom-analyses.	Complete verwerking van piekenlijst uit MS-analyse tot aan statistische analyse.
MASCOT	Gratis Mascot Server (beperkte grootte), licentie nodig voor hoge troughput.	Breed toegepaste analysetool in wetenschappelijk onderzoek.
Comet	Open-source voor Linux en Windows.	Aanvullende softwaretools zijn nodig voor de analyse van de outputdata.
IPSA	Open-source webtool.	Toepasbaarheid voor metaproteoom-data niet bewezen.

Uitdagingen bij de data-interpretatie

Een belangrijke limitatie in metaproteoom-analyses blijft de resolutie van de metaproteoom-datasets. Daarnaast blijven de databases momenteel zeker voor drinkwatermonsters erg beperkt, zoals ook in 3.2.5 werd aangestipt.

Een belangrijke stap binnen metaproteomics-onderzoek op drinkwatermonsters is daarom het gebruik van metagenoom-data als referentie. Voor het genereren van deze referentiedata worden de stappen die beschreven zijn in hoofdstuk 2 toegepast. Uitdagingen voor metagenoom-referentiedata blijven de lage compleetheid van veel metagenome-assembled genomes (MAGs) en de relatief grote aantallen "**hypothetical proteins**" tijdens de annotatie.

3.3 Eerdere metaproteoom-studies aan (drink)water

Er is slechts beperkt onderzoek gedaan naar metaproteomics in drinkwater. Het gepubliceerde onderzoek is beperkt tot aquifers (diepe ondergrondse waterhoudend gesteente/sediment) en verontreinigde grondwaterlagen. In een studie van Benndorf *et al.* (2007) wordt metaproteomics toegepast op grondwatermonsters die verontreinigd zijn met chloorbenzeen (Benndorf et al. 2007). Hierdoor bevat het system tussen de 10^7 en 10^8 microbiële cellen, hetgeen een normaal aantal is voor aquifers. In grondwater bevinden zich meestal veel minder micro-organismen. In deze studie was daarom de filtratie van 1L grondwater (16,000 xg, 10 minuten bij 4°C) en extractie met natriumhydroxide (NaOH) en fenol voldoende. In totaal werden in deze studie 3,808 eiwitgroepen geïdentificeerd

Een studie van Starke *et al.* (2017) heeft gekeken naar het metaproteoom van anoxisch grondwater in een 65

meter diepe aquifer in Hainich, Duitsland (Starke et al. 2017). Hiervoor is in totaal een volume van 1000 liter water gefilterd over een 0.3 µm glasfilter. In de studie is niet gerapporteerd wat de opbrengst van deze extractie was. Extractie was uigevoerd met centrifugatie en als chemicaliën SDS, fenol en ammoniumacetaat.
In een zeer recente studie van Capriotti *et al.* (2021) wordt ook gekeken naar het metaproteoom van een aquifer. Bij de extractie is gebruik gemaakt van spiken (aanenten) met een eiwitextract van gist om de opbrengst van eiwitten uit het complexe monster te verhogen. Het beste resultaat werd hier bereikt bij het gebruik van een tweetal buffers: een buffer met een denaturerende chemicaliën en een buffer met een zuur organisch oplosmiddel. Echter konden in totaal maar 239 eiwitten worden gereconstrueerd, een relatief laag aantal in vergelijk met de studie van bijvoorbeeld Starke et al. waarbij bijna 4 duizend eiwitgroepen werden geïdentificeerd.
Aan oppervlaktewater is meer metaproteoom-onderzoek gedaan, al focussen de meeste studies op mariene ecosystemen en zoutwatermeren. Een studie van Hanson, Hewson & Madsen uit 2014 geeft een overzicht van zes verschillend aquatische habitats, waaronder een tweetal meren en een brakwaterdelta (Hanson, Hewson, and Madsen 2014). Metaproteomics droeg in deze studie bij aan het verkrijgen van inzicht van de dominante energiemetabolismen en rollen in de koolstof-, stikstof- en zwavelcyclus. Specifieke studies aan het meromictische Ace Lake (een meer waarvan de bovenste en onderste waterlaag nooit met elkaar vermengen) in Antarctica hebben daarnaast specifiek gekeken naar bacteriochlorofyl-diversiteit en groei van micro-organismen in zeer nutriëntarme systemen (Lauro et al. 2011; Ng et al. 2010). De kennis over metaproteomics van zoet oppervlaktewater is dus beperkt.

3.4 De meerwaarde en uitdagingen van metaproteomics

Uit dit onderzoek kan geconcludeerd worden dat metaproteomics waardevolle informatie kan geven over de biochemische processen in drinkwaterzuivering en -distributie. Met metaproteomics is het mogelijk om informatie te krijgen over actieve biochemische processen die plaatsvinden in de microbiële populatie. Hiermee kunnen bijvoorbeeld uitspraken gedaan worden over micro-organismen die betrokken zijn bij biofilmvorming. Metaproteomics heeft daarmee een toegevoegde waarde op metagenomics waarmee alleen het metabole potentieel in kaart kan worden gebracht. Wel vormt metagenomics een belangrijke basis voor de data-interpretatie tijdens metaproteomics.

Uit de verkenning van de verschillende methoden en uit gesprekken met proteomics-expert Asst. Prof. dr. Martin Pabst van de Microbial Proteomics Group aan de TU in Delft volgt dat de optimalisatie van metaproteomics voor drinkwatermonsters zeer bewerkelijk en complex is. Aanvullend geldt dat voor de analyse binnenshuis een metaproteomics-infrastructuur dient te worden opgezet, waarbij aanvullend voor de LC-MS-methoden specifieke apparatuur nodig is (nanospray ionisatie voor peptiden) die momenteel niet bij KWR aanwezig is. De praktische toepassing van metaproteomics-experimenten is op korte termijn alleen haalbaar wanneer de samenwerking wordt gezocht met onderzoekslaboratoria die een dergelijke infrastructuur hebben. Binnen Nederland is de Microbial Proteomics Group in Delft gezien de aanwezige faciliteiten en ervaring met drinkwatermonsters hiervoor een goede kandidaat. Eerste verkennende gesprekken hebben plaatsgevonden en er is wederzijdse interesse om een samenwerkingsverband op te zetten. Het is van belang om hiervoor kritisch te kijken hoe beide partijen kunnen profiteren van een dergelijke samenwerking.
Voor KWR en de drinkwaterbedrijven ligt de meerwaarde in dat zonder veel inspanning en kosten de voordelen van proteomics ten opzichte van metagenomics kunnen worden verkend. Met een eventuele samenwerking kan op een betrouwbare manier meer duidelijkheid worden verkregen over de microbiële processen die in de zuivering en distributie plaatsvinden. Voor de Microbial Proteomics Group ligt de meerwaarde voornamelijk in de verbreding van het onderzoek aan drinkwatermonsters, wat kan bijdragen aan lopende onderzoeken en een lopend PhD-project. Het is hier tevens van belang dat goed wordt afgestemd welke delen van de gegenereerde datasets uiteindelijk (wetenschappelijk) gepubliceerd kunnen en mogen worden. De publicatie van data kan uiteindelijk een

meerwaarde hebben voor alle betrokken partijen.
De meerwaarde en uitdagingen zijn hieronder samengevat.

Meerwaarde

- Functionele informatie over de biochemische processen in een microbiële populatie aan de hand van de eiwitsamenstelling.
- Er bestaan goede protocollen voor metaproteoom-onderzoek op drinkwatermonsters, met name bij de Microbial Proteomics Group aan de TU Delft. Samenwerking opzetten met deze groep is veelbelovend.
- Semi-kwantitatieve informatie over de meest abundante biochemische processen in een systeem.
- Biedt in combinatie met metagenoom-data een beeld van het biochemisch potentieel en de actieve processen.

Uitdagingen

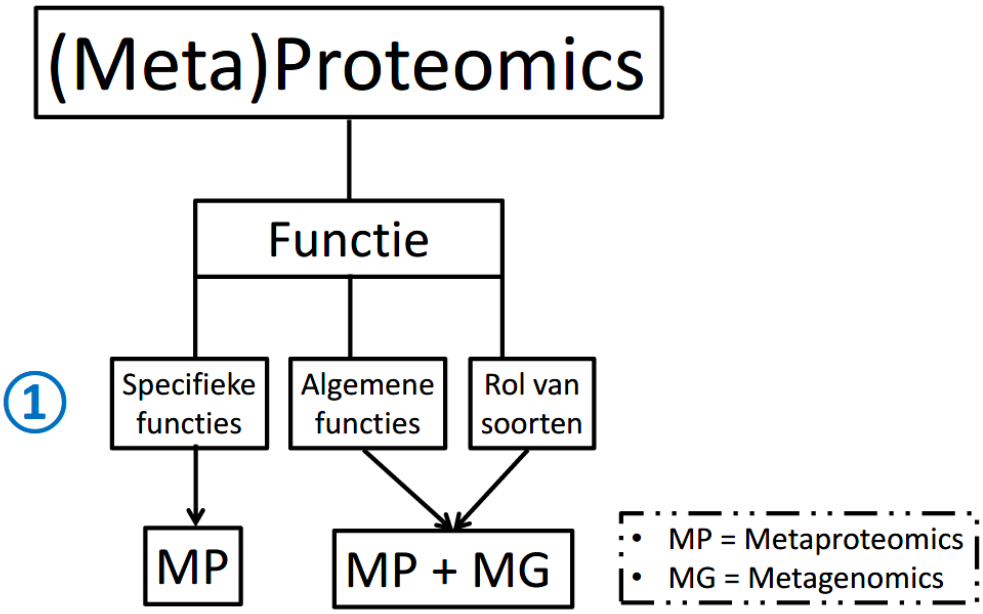
- De resolutie in eiwit-extractie, massaspectrometrische analyse en de identificatie van peptiden (databases) is beperkt.
- Het is niet haalbaar een metaproteoom workflow op korte termijn op te zetten bij KWR door limitaties in materieel en ervaring.
- Eukaryote eiwitten kunnen een storend effect op de resolutie van de bacteriële data hebben en vice versa. Ook kan een lage hoeveelheid eiwitten die verwacht kan worden in (sommige) dinkwatermonsters zorgen voor resolutieproblemen.
- Metaproteomics is vooral qua laboratoriumwerkzaamheden een veel uitgebreidere en dus tragere methode in vergelijking met DNA-gebaseerde technieken.

3.4.1 Implementatiemogelijkheden van metaproteomics

De hoofdvraag van deze bureaustudie binnen het BTO-project Moleculair Microbiologische Methoden betreft de toepassingsmogelijkheden van methoden waarmee de rol van microbiologie bij zuiveringsprocessen en in het distributiesysteem kan worden onderzocht. In dit hoofdstuk is daarbij specifiek naar de potentie van metaproteomics gekeken.
Metaproteomics is een relatief nieuwe techniek die momenteel niet breed binnen het drinkwateronderzoek wordt toegepast. Eerdere verkenningen zijn uitgevoerd, maar hiervoor zijn op dit moment nog geen wetenschappelijke publicaties beschikbaar.
Als onderdeel van deze bureaustudie naar de implementatiemogelijkheden van metaproteomics is allereerst een overzicht samengesteld voor de benodigde stappen van een metaproteoom-analyse. Het doel hiervan is om een beter inzicht in de metaproteoom-workflow te krijgen (Figuur 19). Daarnaast is een inventarisatie gedaan naar de haalbaarheid voor de implementatie van een metaproteomics-workflow. Hieruit bleek dat, gezien de beperkte kennis op het gebied en de noodzaak van aanvullend materieel, het hiervoor het meest haalbaar is om in eerste instantie een samenwerking op te zetten met een instituut dat over deze expertise beschikt. Hierbij kwam in deze verkenning de Microbial Proteomics Group aan de TU in Delft naar voren. Binnen dit onderzoek hebben de eerste verkennende gesprekken plaatsgevonden.
Wanneer blijkt dat metaproteomics een vaste waarde wordt in het drinkwater-onderzoek, kan op de langere termijn gekozen worden om binnen KWR een eigen metaproteomics-infrastructuur op te gaan zetten.

3.5 Flowchart voor de selectie van de juiste methode voor een metaproteomics-project

Om de keuze voor de geschikte aanpak voor een metaproteomics-project te vereenvoudigen is hieronder een flowchart weergegeven met hierin de verschillende opties en afwegingen (Figuur 27). Onder de flowchart is een uitgebreide uitleg van de stappen en mogelijke keuzes gegeven.



Figuur 27. Flowchart voor de methoden-keuze tijdens (meta)proteomics-projecten. De stappen en keuzes worden hieronder verder toegelicht.

Hieronder worden de keuzes in meer detail toegelicht.

① Voor functionele studies kan gekeken worden naar specifieke functies, algemene functies of de rol van individuele soorten in een monster.

-Specifieke functies: wanneer het doel is om specifieke functies in een monster in kaart te brengen kan het voldoende zijn om enkel metaproteomics uit te voeren. Het is hier cruciaal dat de eiwitten van deze specifieke functies voldoende aanwezig zijn in het monster. Een onderzoek naar specifieke functies dient zich daarom te richten op dominante processen.

-Algemene functies: wanneer het doel is om algemene functies in kaart te brengen, is de resolutie van de data een uitdaging. Voor een diepere resolutie van de metaproteomics-data is het daarom aan te bevelen om deze te combineren met een metagenomics sequencing-run.

-Rol van soorten: wanneer het doel is om de specifieke rol van soorten in kaart te brengen dient de metaproteomics-data gecombineerd te worden met uit metagenomics-analyse verkregen MAGs. Hiervoor dient daarom een volledige metagenomics-run inclusief binning te worden uitgevoerd.

Naast microbiële functies kan met metaproteomics ook de microbiële diversiteit in kaart worden gebracht. Voor nu is dit echter niet de focus.
Wanneer het doel van het experiment is om de actieve eiwit-diversiteit in kaart te brengen kan gekozen worden uit drie benaderingen: het in kaart brengen van de taxonomische diversiteit van de eiwitten, de relatieve abundantie

van de eiwitten of de identificatie van microbiële soorten aan de hand van (dominante) eiwitten in het monster. Wanneer de verwachte diversiteit in een systeem erg hoog is, wordt metaproteomics hiervoor echter een uitdaging. De resolutie van metaproteomics is namelijk beperkt en kan zich enkel richten op dominante eiwitten in een monster. Wanneer de onderzoeksvraag zich hierop richt is het uitvoeren van enkel een metaproteomics-analyse voldoende.
Om een breder beeld te krijgen van het metabole potentieel én de diversiteit aan eiwitten in een systeem is het in raadzaam om metaproteomics met metagenomics te combineren. De metagenomics-dataset dient hierbij als basis voor de interpretatie van de metaproteomics-data. Hiermee kan doorgaans een grotere hoeveelheid eiwitten in een monster worden gekarakteriseerd. Dit is noodzakelijk wanneer het doel is om een zo volledig mogelijke diversiteit aan eiwitten in een monster in kaart te brengen.

4 Conclusies en aanbevelingen

4.1 Conclusies

- Metagenomics is een waardevolle tool voor het in kaart brengen van het potentieel van microbiële metabole processen en de microbiële diversiteit in een milieu.

- Met beschikbare open source software-tools is het mogelijk om een eigen pipeline voor metagenomics-analyses op te zetten. Dit vergt echter aanvullende kennis (expertise), opzet en validatie. Deze kennis is binnen een aantal universiteiten (RU, WUR) aanwezig, waardoor samenwerking op dit vlak voor de hand ligt met als doel deze kennis eigen te maken bij KWR wanneer blijkt dat metagenomics waardevol is voor het onderzoek.

- Inhoudelijke kennis op het metagenomics-proces is cruciaal voor accurate monstername en data-analyse om de aanwezige bias tot een minimum te beperken en het maximale uit de data te kunnen halen.

- Met metaproteomics wordt functionele informatie verkregen over de actieve biochemische processen die plaatsvinden in een microbiële populatie van een bepaald ecosysteem. Dit geeft waardevolle informatie wanneer de wens bestaat om de voor een bepaald punt te inventariseren welke actieve processen plaatsvinden.

- De kracht van metaproteomics voor drinkwatersystemen wordt significant verhoogd wanneer de analyse wordt gedaan aan de hand van een eerder verkregen metagenomics referentiedataset.

- Het opzetten van metaproteomics-analyses is zeer complex en vereist gedegen voorkennis. Deze expertise en het benodigde materieel is momenteel niet bij KWR aanwezig, waardoor samenwerking met de Microbial Proteomics Group (dr. Martin Pabst) voor de hand ligt.

4.2 Aanbevelingen

Voor vervolgonderzoek naar het ophelderen van microbiologische processen in de drinkwaterpraktijk is het aan te bevelen om op korte termijn op het gebied van metagenomics en metaproteomics samenwerkingen op te zetten met universitaire afdelingen met expertise op deze gebieden. Het leggen van contacten hiervoor vormt onderdeel van dit BTO-project en zal in het verdere verloop van het project worden gedaan.

5 Referenties

5.1 Wetenschappelijke literatuur

Albertsen, Mads, Philip Hugenholtz, Adam Skarshewski, Kåre L. Nielsen, Gene W. Tyson, and Per H. Nielsen. 2013. 'Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes', *Nature Biotechnology*, 31: 533-38.

Arango-Argoty, Gustavo, Emily Garner, Amy Pruden, Lenwood S. Heath, Peter Vikesland, and Liqing Zhang. 2018. 'DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data', *Microbiome*, 6: 23.

Araujo, Fabricio Almeida, Debmalya Barh, Artur Silva, Luis Guimarães, and Rommel Thiago Juca Ramos. 2018. 'GO FEAT: a rapid web-based functional annotation tool for genomic and transcriptomic data', *Scientific Reports*, 8: 1794.

Aziz, Ramy K., Daniela Bartels, Aaron A. Best, Matthew DeJongh, Terrence Disz, Robert A. Edwards, Kevin Formsma, Svetlana Gerdes, Elizabeth M. Glass, Michael Kubal, Folker Meyer, Gary J. Olsen, Robert Olson, Andrei L. Osterman, Ross A. Overbeek, Leslie K. McNeil, Daniel Paarmann, Tobias Paczian, Bruce Parrello, Gordon D. Pusch, Claudia Reich, Rick Stevens, Olga Vassieva, Veronika Vonstein, Andreas Wilke, and Olga Zagnitko. 2008. 'The RAST Server: Rapid Annotations using Subsystems Technology', *BMC Genomics*, 9: 75.

Benndorf, Dirk, Gerd U. Balcke, Hauke Harms, and Martin von Bergen. 2007. 'Functional metaproteome analysis of protein extracts from contaminated soil and groundwater', *The ISME Journal*, 1: 224-34.

Bowers, Robert M., Nikos C. Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T. B. K. Reddy, Frederik Schulz, Jessica Jarett, Adam R. Rivers, Emiley A. Eloe-Fadrosh, Susannah G. Tringe, Natalia N. Ivanova, Alex Copeland, Alicia Clum, Eric D. Becraft, Rex R. Malmstrom, Bruce Birren, Mircea Podar, Peer Bork, George M. Weinstock, George M. Garrity, Jeremy A. Dodsworth, Shibu Yooseph, Granger Sutton, Frank O. Glöckner, Jack A. Gilbert, William C. Nelson, Steven J. Hallam, Sean P. Jungbluth, Thijs J. G. Ettema, Scott Tighe, Konstantinos T. Konstantinidis, Wen-Tso Liu, Brett J. Baker, Thomas Rattei, Jonathan A. Eisen, Brian Hedlund, Katherine D. McMahon, Noah Fierer, Rob Knight, Rob Finn, Guy Cochrane, Ilene Karsch-Mizrachi, Gene W. Tyson, Christian Rinke, Nikos C. Kyrpides, Lynn Schriml, George M. Garrity, Philip Hugenholtz, Granger Sutton, Pelin Yilmaz, Folker Meyer, Frank O. Glöckner, Jack A. Gilbert, Rob Knight, Rob Finn, Guy Cochrane, Ilene Karsch-Mizrachi, Alla Lapidus, Folker Meyer, Pelin Yilmaz, Donovan H. Parks, A. Murat Eren, Lynn Schriml, Jillian F. Banfield, Philip Hugenholtz, Tanja Woyke, and Consortium The Genome Standards. 2017. 'Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea', *Nature Biotechnology*, 35: 725-31.

Brown, Connor L., Ishi M. Keenum, Dongjuan Dai, Liqing Zhang, Peter J. Vikesland, and Amy Pruden. 2021. 'Critical evaluation of short, long, and hybrid assembly for contextual analysis of antibiotic resistance genes in complex environmental metagenomes', *Scientific Reports*, 11: 3753.

Chao, Yuanqing, Liping Ma, Ying Yang, Feng Ju, Xu-Xiang Zhang, Wei-Min Wu, and Tong Zhang. 2013. 'Metagenomic analysis reveals significant changes of microbial compositions and protective functions during drinking water treatment', *Scientific Reports*, 3: 3550.

Chaudhuri, Roy R., Nicholas J. Loman, Lori A.S. Snyder, Christopher M. Bailey, Dov J. Stekel, and Mark J. Pallen. 2007. 'x BASE2: a comprehensive resource for comparative bacterial genomics', *Nucleic Acids Research*, 36: D543-D46.

Chaumeil, Pierre-Alain, Aaron J Mussig, Philip Hugenholtz, and Donovan H Parks. 2019. 'GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database', *Bioinformatics*, 36: 1925-27.

Ejigu, Girum Fitihamlak, and Jaehee Jung. 2020. 'Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing', *Biology*, 9: 295.

Eng, Jimmy K., Brian C. Searle, Karl R. Clauser, and David L. Tabb. 2011. 'A Face in the Crowd: Recognizing Peptides Through Database Search*', *Molecular & Cellular Proteomics*, 10: R111.009522.

Eveno, Mégane, Yanath Belguesmia, Laurent Bazinet, Frédérique Gancel, Ismail Fliss, and Djamel Drider. 2021. 'In silico analyses of the genomes of three new bacteriocin-producing bacteria isolated from animal's faeces', *Archives of Microbiology*, 203: 205-17.

Gomez-Alvarez, V., R. P. Revetta, and J. W. Santo Domingo. 2012. 'Metagenomic analyses of drinking water receiving different disinfection treatments', *Appl Environ Microbiol*, 78: 6095-102.

Hall, Barry G. 2013. 'Building Phylogenetic Trees from Molecular Data with MEGA', *Molecular Biology and Evolution*, 30: 1229-35.

Hanson, Buck T., Ian Hewson, and Eugene L. Madsen. 2014. 'Metaproteomic Survey of Six Aquatic Habitats: Discovering the Identities of Microbial Populations Active in Biogeochemical Cycling', *Microbial Ecology*, 67: 520-39.

Hyatt, Doug, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. 'Prodigal: prokaryotic gene recognition and translation initiation site identification', *BMC Bioinformatics*, 11: 119.

Jia, Shuyu, Kaiqin Bian, Peng Shi, Lin Ye, and Chang-Hong Liu. 2020. 'Metagenomic profiling of antibiotic resistance genes and their associations with bacterial community during multiple disinfection regimes in a full-scale drinking water treatment plant', *Water Research*, 176: 115721.

Khaleque, Himel Nahreen, Carolina González, D. Barrie Johnson, Anna H. Kaksonen, David S. Holmes, and Elizabeth L. J. Watkin. 2020. 'Genome-based classification of Acidihalobacter prosperus F5 (=DSM 105917=JCM 32255) as Acidihalobacter yilgarnensis sp. nov', *International Journal of Systematic and Evolutionary Microbiology*, 70: 6226-34.

Konstantinidis, K. T., and J. M. Tiedje. 2005. 'Genomic insights that advance the species definition for prokaryotes', *Proc Natl Acad Sci U S A*, 102: 2567-72.

Lauro, F. M., M. Z. DeMaere, S. Yau, M. V. Brown, C. Ng, D. Wilkins, M. J. Raftery, J. A. Gibson, C. Andrews-Pfannkoch, M. Lewis, J. M. Hoffman, T. Thomas, and R. Cavicchioli. 2011. 'An integrative study of a meromictic lake ecosystem in Antarctica', *Isme j*, 5: 879-95.

Leary, Dagmar Hajkova, W. Judson Hervey, Jeffrey R. Deschamps, Anne W. Kusterbeck, and Gary J. Vora. 2013. 'Which metaproteome? The impact of protein extraction bias on metaproteomic analyses', *Molecular and Cellular Probes*, 27: 193-99.

Li, Heng, and Richard Durbin. 2009. 'Fast and accurate short read alignment with Burrows–Wheeler transform', *Bioinformatics*, 25: 1754-60.

Li, Qi, Shuili Yu, Shengfa Yang, Wei Yang, Sisi Que, Wenjie Li, Yu Qin, Weiwei Yu, Hui Jiang, and Deqiang Zhao. 2021. 'Eukaryotic community diversity and pathogenic eukaryotes in a full-scale drinking water treatment plant determined by 18S rRNA and metagenomic sequencing', *Environmental Science and Pollution Research*, 28: 17417-30.

Moriya, Yuki, Masumi Itoh, Shujiro Okuda, Akiyasu C. Yoshizawa, and Minoru Kanehisa. 2007. 'KAAS: an automatic genome annotation and pathway reconstruction server', *Nucleic Acids Research*, 35: W182-W85.

Ng, C., M. Z. DeMaere, T. J. Williams, F. M. Lauro, M. Raftery, J. A. Gibson, C. Andrews-Pfannkoch, M. Lewis, J. M. Hoffman, T. Thomas, and R. Cavicchioli. 2010. 'Metaproteogenomic analysis of a dominant green sulfur bacterium from Ace Lake, Antarctica', *Isme j*, 4: 1002-19.

Oh, Seungdae, Frederik Hammes, and Wen-Tso Liu. 2018. 'Metagenomic characterization of biofilter microbial communities in a full-scale drinking water treatment plant', *Water Research*, 128: 278-85.

Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. 2015. 'CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes', *Genome Research*, 25: 1043-55.

Poghosyan, Lianna, Hanna Koch, Jeroen Frank, Maartje A. H. J. van Kessel, Geert Cremers, Theo van Alen, Mike S. M. Jetten, Huub J. M. Op den Camp, and Sebastian Lücker. 2020. 'Metagenomic profiling of ammonia- and methane-oxidizing microorganisms in two sequential rapid sand filters', *Water Research*, 185: 116288.

Rodriguez-R, Luis M., and Konstantinos T. Konstantinidis. 2016. 'The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes', *PeerJ Preprints*, 4: e1900v1.

Rutherford, Kim, Julian Parkhill, James Crook, Terry Horsnell, Peter Rice, Marie-Adèle Rajandream, and Bart Barrell. 2000. 'Artemis: sequence visualization and annotation', *Bioinformatics*, 16: 944-45.

Seemann, Torsten. 2014. 'Prokka: rapid prokaryotic genome annotation', *Bioinformatics*, 30: 2068-69.

Sieber, Christian M. K., Alexander J. Probst, Allison Sharrar, Brian C. Thomas, Matthias Hess, Susannah G. Tringe, and Jillian F. Banfield. 2018. 'Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy', *Nature Microbiology*, 3: 836-43.

Starke, Robert, Martina Müller, Michael Gaspar, Manja Marz, Kirsten Küsel, Kai Uwe Totsche, Martin von Bergen, and Nico Jehmlich. 2017. 'Candidate Brocadiales dominates C, N and S cycling in anoxic groundwater of a pristine limestone-fracture aquifer', *Journal of Proteomics*, 152: 153-60.

Teeling, Hanno, Anke Meyerdierks, Margarete Bauer, Rudolf Amann, and Frank Oliver Glöckner. 2004. 'Application of tetranucleotide frequencies for the assignment of genomic fragments', *Environmental Microbiology*, 6: 938-47.

Vallenet, D., S. Engelen, D. Mornico, S. Cruveiller, L. Fleury, A. Lajus, Z. Rouy, D. Roche, G. Salvignol, C. Scarpelli, and C. Médigue. 2009. 'MicroScope: a platform for microbial genome annotation and comparative genomics', *Database*, 2009.

Vallenet, David, Alexandra Calteau, Stéphane Cruveiller, Mathieu Gachet, Aurélie Lajus, Adrien Josso, Jonathan Mercier, Alexandre Renaux, Johan Rollin, Zoe Rouy, David Roche, Claude Scarpelli, and Claudine Médigue. 2017. 'MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes', *Nucleic Acids Research*, 45: D517-D28.

van der Walt, Andries Johannes, Marc Warwick van Goethem, Jean-Baptiste Ramond, Thulani Peter Makhalanyane, Oleg Reva, and Don Arthur Cowan. 2017. 'Assembling metagenomes, one community at a time', *BMC Genomics*, 18: 521.

Xu, Ling, Zhaobin Dong, Lu Fang, Yongjiang Luo, Zhaoyuan Wei, Hailong Guo, Guoqing Zhang, Yong Q Gu, Devin Coleman-Derr, Qingyou Xia, and Yi Wang. 2019. 'OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species', *Nucleic Acids Research*, 47: W52-W58.

Yang, Ying, Xiaotao Jiang, Benli Chai, Liping Ma, Bing Li, Anni Zhang, James R. Cole, James M. Tiedje, and Tong Zhang. 2016. 'ARGs-OAP: online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured ARG-database', *Bioinformatics*, 32: 2346-51.

Yuan, Cheng, Jikai Lei, James Cole, and Yanni Sun. 2015. 'Reconstructing 16S rRNA genes in metagenomic data', *Bioinformatics*, 31: i35-i43.

Zhou, Zhenchao, Lan Xu, Lin Zhu, Yang Liu, Xinyi Shuai, Zejun Lin, and Hong Chen. 2021. 'Metagenomic analysis of microbiota and antibiotic resistome in household activated carbon drinking water purifiers', *Environment International*, 148: 106394.

5.2 BTO-rapporten

Heijnen, L. (2018). Toepassingsmogelijkheden van: Metatranscriptomics, Microbial profiling en MinION sequencing. BTO 2018.044. KWR, Nieuwegein.

Heijnen, L. (2019). Verkenning van de toepassingsmogelijkheden van MinION sequencing. BTO 2019.057. KWR, Nieuwegein.

Hornstra, L. (2017). Antibioticaresistentiegenen in oppervlaktewater en in drinkwater zuiveringsprocessen. BTO 2017.042. KWR, Nieuwegein.

Learbuch, K., Timmers, P. (2020). Microbiologische compositie en potentiele metabole processen tijdens drinkwaterdistributie. BTO 2020.044. KWR, Nieuwegein.

Pronk, T., Zwartsen, A., Meekel, N. Dingemans, M. (2021). Functional omics techniques for drinking water quality monitoring. BTO 2021.008. KWR, Nieuwegein.

Segrave, A., Medema, G. (2019). Verkennend onderzoek 2018-2023. BTO 2019.202. KWR, Nieuwegein.

5.3 Nieuwsbrieven

Heijnen, L., Timmers, P. (2019). Nieuwsbrief: kansen van nieuwe moleculair-microbiologische methoden voor toepassing in de drinkwatersector. KWR, Nieuwegein.

5.4 Lijst van websites

ANI/AAI Tool Kostas Lab - Kostas lab | Tools (gatech.edu)

Burrows-Wheeler Alignment (BWA) tool - Burrows-Wheeler Aligner (sourceforge.net)

CheckM – CheckM - assessing the quality of genome bins (ecogenomics.github.io)

CLC Genomics - QIAGEN CLC Genomics - Bioinformatics Software and Services | QIAGEN Digital Insights

DAS Tool - GitHub - cmks/DAS_Tool: DAS Tool

DNA Databank of Japan (DDNJ) – DDBJ (nig.ac.jp)

European Nucleotide Archive (ENA) – ENA Browser (ebi.ac.uk)

Genbank – GenBank Overview (nih.gov)

GO FEAT - GO FEAT | fast functional annotation web tool for genomic and transcriptomic data (ufpa.br)

GTDB-Tk - GitHub - Ecogenomics/GTDBTk: GTDB-Tk: a toolkit for assigning objective taxonomic classifications to bacterial and archaeal genomes.

Hidden Markov Models (HMMs) - A.pdf (stanford.edu)

hybridSPAdes - GitHub - ablab/spades: SPAdes Genome Assembler

IDBA-UD - GitHub - loneknightpy/idba

InterPro - InterPro (ebi.ac.uk)

Kaiju - Kaiju: Fast and sensitive taxonomic classification for metagenomics (ku.dk)

KEGG Automated Annotation Server (KAAS) - KAAS - KEGG Automatic Annotation Server (genome.jp)

MEGA - Home (megasoftware.net)

MEGAHIT - GitHub - voutcn/megahit: Ultra-fast and memory-efficient (meta-)genome assembler

MEGAHIT workshop - MEGAHIT Assembly — Metagenomics Workshop 1 documentation (denbi-metagenomics-workshop.readthedocs.io)

MetaCyc - MetaCyc: Metabolic Pathways From all Domains of Life

Metaspades - metaSPAdes – Center for Algorithmic Biotechnology (spbu.ru)

MG-RAST - MG-RAST

MicroScope - MicroScope Home - MaGe: Microbial Genome Annotation & Analysis Platform - MicroScope - Web Interface System & Specialized Databases for (re)Annotation and Analysis of Microbial Genomes (cns.fr)

NCBI Prokaryotic Genomes Automatic Annotation Pipeline - NCBI Prokaryotic Genome Annotation Pipeline (nih.gov)

OPERA-MS - GitHub - CSB5/OPERA-MS: OPERA-MS - Hybrid Metagenomic Assembler

Prodigal - compbio.ornl.gov

Prokka - GitHub - tseemann/prokka: Rapid prokaryotic genome annotation

Protter - Protter - interactive protein feature visualization (ethz.ch)

RAST - RAST Server - RAST Annotation Server (nmpdr.org)

Spades - SPAdes – Center for Algorithmic Biotechnology (spbu.ru)

Taxonomer – taxonomer.iobio Sample Datasets

TMHMM Server - TMHMM Server, v. 2.0 (dtu.dk)

UniProt – UniProt

Sequence-Based Ultra-Rapid Pathogen Identification (SURPI) - SURPI

xBASE2 - xbase.bham.ac.uk

I Begrippenlijst

I.I Begrippenlijst Hoofdstuk 2 - Metagenomics

16S rRNA gen	Het 16S rRNA komt voor in prokaryoten (bacteriën en archaea) en wordt gecodeerd door het 16S rRNA gen. Het vormt onderdeel van het ribosoom, het eiwit-synthetiserend organel van een cel
AAI	Afkorting van 'average amino acid identity', ofwel de gemiddelde identiteit tussen twee aminozuursequenties, uitgedrukt in een percentage. Meestal wordt dit gebruikt om gehele genomes met elkaar te vergelijken.
Abundantie	De hoeveelheid/aanwezigheid van een bepaald gen, micro-organisme of bijvoorbeeld een bepaalde groep micro-organismen in een monster. Vaak wordt abundantie bepaald aan de hand van de hoeveelheid DNA-reads in een monster. Dit is een benadering voor de daadwerkelijke abundantie van een micro-organisme.
Actieve site	De locatie in een eiwit/enzym dat de chemische reactie uitvoert.
Alignment	Het proces waarbij DNA-sequenties van eenzelfde gen of gen-regio onder elkaar worden gelegd en worden overlapt op dezelfde gebieden. Dit vormt de basis voor het berekenen van fylogenie en fylogenetische bomen.
Aminozuursequentie	De sequentievolgorde van eiwitten vanuit de bouwstenen, de aminozuren. Er zijn 20 basis-aminozuren en 2 aanvullende aminozuren (selenocysteine en pyrrolysine) die door modificaties zijn ontstaan en voorkomen in specifieke micro-organismen
Amplicon sequencing	Het proces waarbij een groot aantal kopieën van eenzelfde DNA-fragment wordt gegenereerd via PCR en vervolgens wordt gesequenced. Een goed voorbeeld hiervan is 16S rRNA sequencing waarbij (meestal) een kort fragment van het 16S rRNA gen wordt geselecteerd, geamplificeerd met PCR en vervolgens wordt gesequenced. Er kunnen ook andere (functionele) genen worden geselecteerd voor amplicon sequencing
ANI	Average Nucleotide Identity. Een percentage van de gemiddelde nucleotide-identiteit tussen genomes van verschillende micro-organismen. Dit wordt gebruikt als een maat voor de identiteit van verschillende micro-organismen.
Annotatie	Het proces waarbij met behulp van een referentiedatabase in DNA sequencing-data wordt gezocht naar de aanwezigheid en identiteit van genen. Hier wordt in de regel de geassembleerde (aan elkaar geplakte) DNA-data of zelfs de MAGs voor gebruikt.
Archaea / archaeal	Behorend tot het domein van de Archaea.

Assemblage / assembly	Het process waarbij korte DNA-fragmenten aan elkaar worden geplakt om grotere fragmenten, contigs geheten, te vormen.
Bacterie / bacterieel	Behorend tot het domein van Bacteriën.
Base calling	Het proces bij methoden van Oxford Nanopore waarbij met software de nanoporie-signalen worden omgezet naar nucleotiden-informatie.
Bias	De vertekening van de data door afwijkingen in de gebruikte methode(n). Bij DNA-sequencing zijn bekende punten waarbij biases optreden de DNA-extractie en de DNA-sequencing.
Binning	Het proces waarbij vanuit geassembleerde (aan elkaar geplakte) DNA-data genomes worden gereconstrueerd. Deze gereconstrueerde genomes worden in de regel bins of metagenome assembled genomes (MAGs) genoemd.
Biologische functie	De functie die bijvoorbeeld een eiwit van nature vervult.
Biotische factor	Een invloed die door een levend organisme op een andere wordt uitgeoefend.
Centraal energiemetabolisme	Het basis energiemetabolisme van een cel of organisme.
Chimere contigs	Lange aan elkaar geknoopte DNA-sequenties die sterk op elkaar lijken, waardoor de verwerking vaak problemen oplevert voor de gebruikte software.
Compleetheid / completeness	Term die bij CheckM wordt gebruikt om aan de hand van een serie van 43 markergenen die één keer in een genoom voorkomen de compleetheid van metagenoom-geassembleerde genomes, ofwel MAGs te beoordelen. Voor compleetheid geldt dat bij 100% alle markergenen volledig aanwezig zijn.
Contaminatie / contamination	Term die bij CheckM wordt gebruikt om aan de hand van een serie van 43 markergenen die één keer in een genoom voorkomen DNA-data van verschillende organismen in metagenoom-geassembleerde genomes, ofwel MAGs zitten. Voor contaminatie geldt dat de twee kopieën van het markergen een aminozuur-identiteit hebben van minder dan 90%.
Contig	Een lange, door software aaneengeknoopte DNA-sequentie die verkregen is door kortere DNA-fragmenten die uit een sequencing-platform rollen.
Cutoff / drempelwaarde	Een grenswaarde waaronder sequenties of delen van sequenties niet meegenomen worden in de analyse
Dekking / coverage	De mate waarin bepaalde DNA-fragmenten van eenzelfde organisme voorkomen in een metagenoom-dataset.

Detection-based approach	Methode waarbij het doel om met hoge precisie, maar relatief lage gevoeligheid de aanwezigheid van specifieke organismen vast te stellen. Vaak gaat het hierbij om pathogenen.
DNA-sequencing	Het proces waarbij van DNA-moleculen nucleotide-data wordt verkregen.
EC-nummer	Het specifieke nummer dat een eiwit krijgt voor de EC-database. De codering geeft informatie over de categorie en functie van het eiwit die kan worden teruggevonden in de EC-database.
Ecologisch netwerk	Een weergave van de biotische (ofwel “levende”) interacties in een ecosysteem. Dit laat zien hoe de verschillende soorten in een ecosysteem met elkaar verbonden zijn en wat voor interacties deze soorten aangaan. Dit zijn trofische interacties (de ene soort is voedsel voor de andere soort) of bijvoorbeeld symbiotische interacties waarbij beide soorten elkaar helpen en versterken.
Eiwitdomein	Een geconserveerd onderdeel van een eiwit met een specifieke ruimtelijke structuur die zich zelfstandig ontwikkelt (evolueert) en zelfstandig functioneert. Een gemiddeld eiwitdomein is samengesteld uit 50 tot 250 aminozuren.
Eiwitfamilie	Een groep van eiwitten met ongeveer dezelfde bouw. In de meeste gevallen is het actief centrum van deze eiwitten gelijk en hebben de eiwitten eenzelfde of zeer vergelijkbare functie.
Energiemetabolisme	De processen die bijdragen aan de energiehuishouding van een organisme.
Eukaryoot	Behorend tot het domein van de eukaryoten. Eukaryoten zijn organismen waarvan iedere cel een celkern bevat. Het eukaryote DNA onderscheidt zich van het prokaryote DNA door onder andere de aanwezigheid van het 18S rRNA gen, daar waar prokaryoten het 16S rRNA gen bezitten.
Extracellulair	Aanwezig buiten de cel.
Familie (taxonomie)	Taxonomische rang na soort en geslacht (genus).
Fenol	Ook wel hydroxybenzeen genoemd. Chemische stof bestaande uit een benzeenring met een OH-groep. Fenol is een aromatische alcohol die veel wordt toegepast in de extractie van DNA, RNA en eiwitten.
Functionele genen	Genen die coderen voor eiwitten met een bepaalde functie in een organisme.
Fylogenetische boom	Grafische weergave van het evolutionaire verband tussen verschillende organismen.
Fylogenie	De studie van het evolutionaire verband tussen organismen
Fylum	Taxonomisch niveau na klasse en vóór rijk.

GC-percentage	Het percentage van de nucleotiden G en C ten opzichte van T en A. Dit wordt gebruikt als vingerafdruk voor micro-organismen en als dimensie voor binning.
Gencluster	Een groep van twee of meer genen die in een cluster op het DNA worden aangetroffen. Deze genen coderen voor vergelijkbare eiwitten die een gezamenlijke functie hebben.
Genus (taxonomie)	Taxonomisch niveau na soort en vóór familie.
Grafische gebruikersomgeving / GUI	Een interactieve omgeving op een computer waarbij een gebruiker met muisklikken commando's kan geven en analyses kan uitvoeren.
Heterogeniteit / heterogeneity	Term die bij CheckM wordt gebruikt om aan de hand van een serie van 43 markergenen die één keer in een genoom voorkomen DNA-data van twee verschillende organismen in metagenoom-geassembleerde genomes, ofwel MAGs zitten. Voor heterogeniteit geldt dat de twee kopieën van het markergen een aminozuur-identiteit hebben van meer dan 90%.
HMMs	De afkorting voor Hidden Markov Models. Statistische modellen waarmee een proces met onbekende parameters kan worden gemodelleerd. Dit wordt bijvoorbeeld veel gebruikt in het voorspellen of detecteren van verschillende genen/eiwitten met naar verwachting eenzelfde of een vergelijkbare functie.
Hypothetical protein	Een gen dat codeert voor een eiwit waarvan de functie niet bekend is, maar waarvan door de samenstelling wel wordt aangenomen dat wel een specifieke functie aanwezig is.
Illumina	Producent van sequencing-technologie gebaseerd op fluorescente lichtdetectie door fluorescente nucleotiden.
Intracellulair	Aanwezig in een cel.
K-mers	Stukken DNA met een lengte van k nucleotide die gebruikt worden voor de clustering en vorming van contigs, ofwel lange aaneengesloten stukken DNA-sequenties.
Leesframe	De afleesmethode van trinucleotiden (combinaties van 3 nucleotiden die coderen voor een aminozuur) in een DNA-sequentie. Door de mogelijkheid om het leesframe met 1 of 2 nucleotiden naar links of rechts te verplaatsen, zijn er verschillende leesframes en dus ook verschillende vertalingen naar aminozuren mogelijk.
Long-read technologie	Een DNA sequencing-technologie waarbij lange stukken DNA in één keer gesequenced kunnen worden. Deze stukken zijn vaak minimaal enkele duizenden basenparen lang. Bijvoorbeeld bij PacBio zijn gemiddelde DNA-strengen tot 5,000 basenparen en bij de Oxford Nanopore MinION zelfs tot meer dan 100,000 basenparen

Manuele curering	Het proces waarbij handmatig een DNA-sequentie die door een softwareprogramma is gealigned (korte DNA sequenties die aan elkaar worden geplakt) en geannoteerd (het proces waarbij genen en overige coderende sequenties worden geidentificeerd) wordt gecontroleerd, gecorrigeerd en aangevuld.
Messenger RNA (mRNA)	Het RNA dat codeert voor enzymen/eiwitten. Het messenger RNA is de tussenstap tussen het DNA en het eiwit.
Metabole potentieel	De mogelijke functies die een organisme op basis van de op het DNA geïdentificeerde genen kan uitvoeren in een milieu. Dit zijn dus niet de processen die ook daadwerkelijk uitgevoerd worden.
Metagenoom sequencing	Ook wel “metagenomics”genoemd. Het proces waarbij het volledige DNA van een monster wordt gesequenced/gekarakteriseerd.
Metaproteoom sequencing	Ook wel “metaproteomics” genoemd. Het proces waarbij de volledige eiwitsamenstelling van een monster wordt gesequenced/gekarakteriseerd.
MinION	Sequencer voor het sequencen van lange DNA-sequenties. De sequencer is geproduceerd door Oxford Nanopore.
N50	De mediaan van de geassembleerde DNA-sequenties uit een sequencing-analyse. Dit wordt gebruikt als een maat voor het beoordelen, controleren en vergelijken van de assemblage.
NCBI BLAST nr	Database die wordt gebruikt voor het vergelijken van de in een onderzoek verkregen DNA-sequenties (meestal gen-sequenties). De database is samenstesteld door het National Center for Biotechnology Information (NCBI) en bestaat uit een niet redundante (nr) versie, hetgeen betekent dat dubbele sequenties uit de database zijn verwijderd om analyses te verbeteren en te versnellen.
Next-Generation Sequencing	Ofwel afgekort ‘NGS’, een verzamelnaam voor sequencing-methoden ontwikkeld na Sanger sequencing. In de praktijk vaak gebruikt om een amplicon-sequencing studie te beschrijven op een NGS-machine.
Noncoding DNA	DNA dat niet codeert voor eiwitten. Dit DNA speelt onder andere een rol in regulatieprocessen. Voor een groot deel van het noncoding DNA is de functie nog onbekend.
Open-source	Openbaar beschikbaar en vaak ook openbaar aanpasbaar. Dit is het geval voor veel bioinformatische software waarvan ontwikkeling plaatsvindt door een grote community (gemeenschap) van onderzoekers.
Orde (taxonomie)	Taxonomisch niveau na familie en vóór klasse.

Ortholoog gen	Genen die gerelateerd aan elkaar zijn door verticale afstemming van eenzelfde voorouder. Orthologe genen coderen voor eiwitten met eenzelfde functie in verschillende soorten.
PacBio	Sequencing-platform waarbij sequencing van DNA strengen zonder amplificatie gedaan kan worden in zeer kleine optische welletjes met lichtdetectie.
Paraloog gen	Homologe genen die door duplicatie in het genoom van eenzelfde organisme zijn geevolueerd tot genen die coderen voor eiwitten met een vergelijkbare, maar niet dezelfde functie.
Pathway	Ook wel reactiepad genoemd. Een serie van interacties van moleculen/eiwitten in een cel waarmee een bepaald product wordt geproduceerd of waarmee een bepaald proces wordt gereguleerd. Een voorbeeld hiervan is de glycolyse.
Pseudogen	Een gen dat een zeer gelijkende DNA-structuur als genen heeft, maar niet functioneert als gen.
qPCR	Kwantitatieve PCR-methode waarbij met behulp van de productie van een fluorescent signaal tijdens het PCR-proces een kwantificatie van een specifiek gen kan worden gedaan.
Q-score / kwaliteitsscore	Kwaliteitsscore waarmee de accuraatheid van DNA-sequencing in kaart wordt gebracht. Een hogere Q-score betekent hierbij een lagere foutenmarge.
RAM-geheugen	Het voor rekenprocessen beschikbare werkgeheugen voor tijdelijke opslag van data. Voldoende RAM-geheugen (vaak meer dan 64 GB) is noodzakelijk voor de verwerking van NGS-data. Servers bieden significant meer RAM-geheugen dan laptops en desktops.
Reactiemechanisme	Het hoe en waarom een bepaalde reactie onder bepaalde condities plaatsvindt. Dit wordt bijvoorbeeld gebruikt voor het in kaart brengen van metabole reacties in een organisme.
RED-waarde	Een waarde die de relatieve evolutionaire afstand (‘relative evolutionary distance’) tussen twee genomes aangeeft. Vaak wordt voor de RED-waarde een schaal tussen de 0 en 1 gehanteerd met 1: de hoogst aangetroffen evolutionaire afstand tussen twee MAGs in de onderzochte dataset.
Referentiedatabase	Een databank met daarin bijvoorbeeld DNA-sequenties of eiwitsequenties die gebruikt worden om DNA- of eiwitsequenties van een organisme te identificeren.
RefSeq	Een openbare databank met geannoteerde DNA-sequenties en eiwitten. De databank wordt door het NCBI beheerd. De databank is niet redundant, hetgeen betekent dat elke sequentie maar één keer voorkomt.

Regulatoir RNA	RNA dat niet codeert voor enzyme/eiwitten, maar een actieve regulatoire rol in de cel vervult
Rekencapaciteit	De hoeveelheid rekenvermogen dat een computer of server heeft, vaak in verband met het RAM-geheugen. Dit is van belang om te bepalen of bioinformatische analyses op een betreffend platform kunnen worden uitgevoerd.
Resolutie	De diepgang van een analyse die aangeeft in hoeveel detail de resultaten bekeken kunnen worden. Een hogere resolutie betekent meer detail in de geanalyseerde data. Dit heeft grotendeels te maken met de hoeveelheid en kwaliteit van de ingaande data.
Ribosomale sequentie	De sequentie van een ribosomaal RNA-gen, zoals bijvoorbeeld het 16S rRNA gen. Hierbij gaat het om een RNA-sequentie, aangezien het RNA in het ribosoom aanwezig is.
Secundair metabool proces	Metabole processen die geen onderdeel uitmaken van het centrale energiemetabolisme. Een voorbeeld van een secundair metabool proces is bijvoorbeeld de productie van afweerstoffen tegen vraat.
Selectieve PCR	Een PCR-reactie waarbij nauwkeurig een enkel of enkele PCR-producten worden gegenereerd. Dit wordt vaak gedaan door gebruik van nauwkeurigere polymerase-enzymen en DNA-blokkeersequenties die zorgen dat gerelateerde maar niet gewenste stukken DNA niet geamplificeerd worden.
Short-read technologie	Sequencing-technologie waarbij slechts korte fragmenten (vaak tot maximaal 600 basenparen) worden gesequenced. Voorbeelden van short-read technologieën zijn Illumina en IonTorrent.
Signaal (eiwit)	Karakteristieke aminozuursequentie die aangeeft dat een eiwit waarschijnlijk van nature binnen of buiten de cel voorkomt.
SILVA database	Een database samengesteld door ARB. Het is een ribosomale RNA-database die manueel gecureerd is. Er zijn zowel databases beschikbaar voor het small subunit (SSU) RNA, ofwel het 16S en 18S rRNA gen en de large subunit (LSU) RNA's, ofwel het 23S en 28S rRNA gen.
Soort (taxonomie)	Het laagste officiële taxonomische niveau. Soorten kunnen daarnaast nog onderverdeeld worden in rassen. Boven soortniveau komen de genera.
Taxonomische compositie	De taxonomische samenstelling van een monster.
Tetranucleotide frequentie	De hoeveelheden van bepaalde stukken DNA van 4 nucleotiden in een DNA-fragment, vaak gebruikt om reads te groeperen en langere DNA-sequenties (contigs) te vormen.
Transcriptie	Het proces waarbij een DNA-sequentie wordt omgezet in een RNA-sequentie.

Translatie	Het proces waarbij een RNA-sequentie wordt omgezet in een eiwit.
Transposon	Een in het genoom verspringend DNA-element. Dit element kan er voor zorgen dat bepaalde genen wel functioneel zijn (transposon niet aanwezig) of niet functioneel zijn (transposon in het gen aanwezig). Transposons zijn hiermee een regulatoir element in het genoom.
TrEMBL	Een onderdeel van de UniProtKB database. Een completere eiwitdatabase dan de manueel gecureerde UniProt database, maar daarmee minder nauwkeurig. De database wordt softwarematig samengesteld en bijgehouden.
Trimming	Proces waarbij bepaalde stukken van gesequencete DNA-moleculen worden afgeknipt doordat deze niet aan de kwaliteitseisen voldoen.
UniProtKB / Swiss-Prot	Manueel gecureerde eiwitdatabase (UniProtKB) die door Swiss-Prot wordt bijgehouden.
Workflow	Een werkschema waarop een proces staat weergegeven. Dit wordt veel gebruikt voor complexe processen van meerdere stappen, zoals bijvoorbeeld het geval is bij metagenomics.
Wrapper tool	Een tool die gebruikt wordt om resultaten die met verschillende programma's zijn gemaakt te verzamelen en samen te vatten. Ook kan een wrapper tool redundante resultaten verwijderen/samenvoegen. Das Tool is een voorbeeld van een wrapper tool die gebruikt wordt voor consensus binning.

I.II Begrippenlijst Hoofdstuk 3 - Metaproteomics

Affiniteit	De mate waarin bijvoorbeeld een eiwit aan een kolom die voor vloeistofscheiding wordt gebruikt kan binden. Een hoge affiniteit betekent dat het eiwit goed bindt, een lage affiniteit betekent dat het eiwit slecht bindt.
Apolair	Waterafstotend of hydrofoob. Geeft aan dat een stof niet of zeer slecht in water oplost of waterafstotend werkt.
Bead-beating	Proces waarbij gebruik wordt gemaakt van kleine pareltjes/deeltjes van bijvoorbeeld glas of metaal om (meestal) biologisch materiaal open te breken en te homogeniseren.
Contaminatie	Besmetting van een monster met materiaal van buitenaf dat de analyse kan verstoren.
Coomassie blue	Blauwe kleurstof die veel gebruikt wordt voor het aankleuren van eiwitten in een polyacrylamide-gel.
Cytoplasmatisch eiwit	Een eiwit dat zich in de celvloeistof (het cytoplasma) bevindt.
de novo	Bij eiwitidentificatie wordt bij een *de novo* benadering de analyse uitgevoerd zonder enige voorkennis over de identiteit van de verschillende eiwitten.
Desorptie	Het proces waarbij een stof loslaat van bijvoorbeeld een chromatografische kolom.
Detergentia	Oplosmiddelen die gebruikt worden om bijvoorbeeld celmembranen in een biologisch monster op te lossen.
Differentiële centrifugatie	Door gebruik te maken van verschillende centrifugatietijden en -stappen kan een complex eiwitmengsel uit elkaar getrokken worden.
Eiwitcomplex	Een groep van eiwitten die door ladingen en chemische interacties die van nature voorkomen of door de eiwit-extractie aan elkaar binden. Eiwitcomplexen zijn door hun grootte vaak niet te scheiden met de SDS-PAGE methode.
Electrospray-ionisatie (ESI)	Proces waarbij een electrospray-bron wordt gebruikt om peptiden te ioniseren, zodat deze vervolgens geanalyseerd kunnen worden met de massaspectrometer.
French press	Ook wel de “cell press” genoemd. Een apparaat waarmee onder hoge druk een vloeistof door een kleine opening geperst wordt, waardoor cellen lyseren en de celinhoud met eiwitten in oplossing komen.
Homogenisatie	Het proces waarmee bijvoorbeeld deeltjes die niet of slecht oplossen in water gelijkmatig worden verdeeld over de vloeistof.
Hypothetical protein	Een voorspelde eiwitsequentie waar vanuit de database geen mogelijke rol kan worden vastgesteld.

in silico	Proces waarbij op een computer een eiwitdigestie wordt uitgevoerd op metagenoom-data om zo een beeld te krijgen van de mogelijke peptiden die in hetzelfde monster dat ook met metaproteomics is geanalyseerd aanwezig zijn.
Liquid chromatography (LC)	De scheiding van het peptidenmengsel met vloeistofchromatografie, het proces waarbij een eiwitmengsel in vloeistof wordt gescheiden door unieke interacties met een kolom.
Mass spectrometry (MS)	Analyse van de gescheiden peptidenmengsels met een massaspectrometer.
Mass to charge ratio (m/z)	De ratio tussen de massa en de lading van een gemeten molecuul (bijvoorbeeld een peptide). Deze ratio is vaak kenmerkend voor bepaalde stoffen en stofdeeltjes en kan hiermee gebruikt worden voor de identificatie van stoffen.
Massaspectrogram	De uitkomst van massaspectrometrische analyse. Hierin wordt vaak gebruik gemaakt van de mass to charge ratio (m/z) en de verblijftijd op de LC-kolom als dimensies.
Membraangebonden eiwit	Een eiwit dat van nature in of gebonden aan de celmembraan voorkomt.
Natieve conformatie	De natuurlijke staat en structuur waarin eiwitten voorkomen.
Peptide fingerprinting	Het proces waarbij de verkregen massa's uit de massaspectrometrische analyse vergeleken met de data uit een grote database. Zo kunnen de peptiden worden geïdentificeerd en omgezet worden naar eiwitsequenties.
Peptide spectral library	Een handmatig verzamelde databank van peptidenspectra voor analyse van verkregen spectrometrie-data.
Peptide-centric	Methode waarbij peptiden niet chemisch of radioactief gelabeld worden voor de analyse. Deze methode is vereist voor een metaproteomics-analyse.
Polair	Waterminnend of hydrofiel. Geeft aan dat een stof goed in water oplost.
Precipiteren	Het neerslaan van eiwitten door het gebruik van chemicaliën als zouten en alcoholen
Reverse phase (RP)	Modus van de chromatografie waarbij een waterafstotende vaste fase en een wateroplosbare vloeistoffase wordt gebruikt. Dit is de standaard chromatografische methoden voor het LC-gedeelte van de LC-MS.
SDS-PAGE	Methode waarbij het eiwitmengsel eerst wordt gedenatureerd in sodium dodecylsulfaat bij 70°C en vervolgens wordt gescheiden op een polyacrylamide (PA) met gel-electroforese (GE), ofwel PAGE.
Sonicatie	Het proces waarbij gebruik wordt gemaakt van geluidsgolven van ultrasone frequentie (>20 kHz))om bijvoorbeeld cellen en celmateriaal open te breken.
Sorptie	Het proces waarbij een stof bindt aan bijvoorbeeld een chromatografische kolom.

Spiken	Proces waarbij een bekende stof (bijvoorbeeld een eiwit of eiwitmengsel) aan een monster wordt toegevoegd om bijvoorbeeld het rendement of de nauwkeurigheid van de gebruikte methode vast te kunnen stellen.
Tandem MS / MS/MS / MS^2	Analysemethode waarbij twee massaspectrometers achter elkaar gekoppeld worden voor een hogere resolutie (zelfde MS-methode) of voor bredere analyse (verschillende MS-methoden).
Tryptische digestie	Het met een enzym opknippen van eiwitten tot kleine eiwitstukjes van maximaal 50 aminozuren. Deze stukjes kunnen vervolgens worden geanalyseerd