Plan van Aanpak

Afstudeerstage

Hoen, Winny 't

Hogeschool Leiden
Opleiding Bio-informatica

Winny ’t Hoen
S1101573

Afstudeerbegeleider:
Tom Kormelink Groot

Naturalis Biodiversity Center
ARISE-project

Stagebegeleider:
Rutger Vos

29 september 2022
Periode 1
Versie 1

Inhoudsopgave

Inleiding

ARISE-project

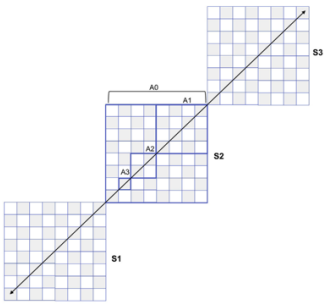
Naturalis lijdt het grote ARISE-project[1]. Het doel van dit project is om alle meercellige soorten in Nederland in kaart te brengen door het bouwen van een soort infrastructuur. Door het bouwen van een infrastructuur kan duidelijk worden wat voor soorten er zijn, waar die soorten zich voornamelijk bevinden en welke relaties er zijn tussen soorten en hun omgeving. Om dit te doen kijken ze naar drie belangrijke voorwaarden; beeld, geluid en DNA. Onderdeel van deze meercellige soorten die worden onderzocht zijn bodemschimmels, waar we tijdens dit project op zullen focussen. Er is *metabarcoding* data beschikbaar van bodemschimmels door het ARISE-project. Deze data waar wij mee zullen werken komt niet van een vreemde plek. De bodemschimmels waar we mee gaan werken zijn verzameld op drie plekken in of vlakbij Leiden, waar Naturalis gevestigd is.

Metabarcoding data

Metabarcoding is een specifiek en uitgebreide techniek voor het identificeren van soorten aan de hand van hun DNA[2]. Dit is ideaal om te gebruiken gezien er ontzettend veel soorten bodemschimmels aanwezig kunnen zijn in de opgehaalde monsters. Metabarcoding geeft ook de mogelijkheid ongeïdentificeerde soorten te identificeren.

Voorafgaand bodemschimmel onderzoek

Afgelopen jaar is er een project gestart om bodemschimmels te verzamelen van drie verschillende locaties; de duinen, een park en een natuur gras gebied[3]. De bemonstering van deze locaties is gebaseerd om een analyse van *Gavito*[4]. De buizen die worden gebruikt om de schimmels uit de grond te halen zijn in een soort schaakbord design, die zich drie keer herhaalt, geplaatst zoals te zien in figuur 1. Vervolgens zijn de bodemschimmels op de volgende manier in het lab geanalyseerd door een student. Eerst werd DNA uit de samples verzamelend, waarna primers werden toegevoegd. Na het toevoegen van primers werd er een PCR uitgevoerd op de bodemschimmels. De laatste pool van de PCR werd overgenomen door Baseclear B.V, Leiden, zij hebben de sequencing uitgevoerd. Vervolgens is de DADA2 ITS pipeline workflow gebruikt om een ampliconsequencievariant (ASV) te ontwikkelen.
Groepen schimmels zijn geïdentificeerd per locatie en naast elkaar gezet in een beta-diversity analyse plot en een alfa analyse plot. Alle samples die op elkaar lijken zijn hierbij geclusterd, afhankelijk van hun locatie. In totaal zijn er 2.581.549 reads gesequenced, waarvan 2.112.745 reads behoren tot bodemschimmels. 82.094 reads konden niet geïdentificeerd worden. Er valt nog veel meer informatie te halen uit de opgehaalde data, dat is waar we het komende project op zullen verder gaan.



Figuur 1: Design voor het bemonsteren van de locaties.

Doelstelling

Er is nieuwe data binnengekomen op Naturalis van een NovaSeq illumina platform. Deze data moeten op dezelfde manier wordt onderzocht als het onderzoek op de bodemschimmels in Leiden. Vervolgens willen wij onderzoeken is of de ruimtelijke schaal van informatie die zijn verzameld van de bodemschimmels genoeg is om de schimmels goed te kunnen onderscheiden en bestuderen. Biedt bijvoorbeeld een van de drie vlakken van het bemonsteren per locatie voldoende informatie, of zelfs nog te weinig informatie. Deze interesse is niet voor de nieuwe data, maar voor de data uit Leiden. Het onderzoek zal dus beginnen bij het her creëren van de DADA2 ITS pipeline workflow op een andere data, om vervolgens het vervolgonderzoek toe te passen op de al bekende data uit Leiden. Om deze doelstelling te behalen is het project op te verdelen in aparte delen. Deze zullen in aparte scripts worden aangeleverd samen met het onderzoeksverslag.

Deel 1

Ampliconsequencievariant (ASV) tabel

Om de ASV-tabel te produceren wordt de DADA2 ITS workflow gevolgd[5]. De illumina gesequenced paired end data verkregen door een NovaSeq illumina platform wordt als input gebruikt van de workflow. Een van de belangrijkste onderdelen van de workflow is het verwijderen van primers. Uiteindelijk wordt een ASV-tabel ontwikkeld. Dit is een uitgebreidere versie van het OTU-tabel, die extra informatie biedt over hoe vaak een bepaalde sequentie variant is aangetroffen in elk monster. Met dit tabel kan verder onderzoek worden uitgevoerd.

Deel 2

Rarefaction curve

Met een rarefaction curve kunnen we bepalen of een bepaalde sample genoeg is gesequenced om zijn identiteit te bepalen[6]. Deze informatie is nodig om de samples te kunnen classificeren. Om een rarefaction curve in elkaar te zetten wordt eerste bekeken hoe vaak elke sample is gesequenced. Deze aantallen kunnen niet van elkaar verschillen, anders kan je geen diversiteit bepalen. Simpel gezegd wordt het aantal samples tegenover de verschillende soorten geplot. Er zal dus wat data verloren gaan tijdens het maken van een rarefaction curve. Tijdens dit proces kan bepaald worden om een aantal samples uit de data te verwijderen, als deze te weinig zijn gesequenced en zorgen voor nog meer onnodig data verlies.

Classificeren van de verschillen tussen de samples

De vervolg stap is het classificeren van de data en laten zien waar de verschillen liggen. Om deze verschillen te kunnen bepalen kunnen meerdere statistische toetsen worden uitgevoerd. Enkele voorbeelden van deze toetsen die gebruikt kunnen worden tijdens dit onderzoek zijn de *Unifrac distanc,* de *Bray-Curtis distance, de Mantel test* en de *species accumulation curve.* De *Unifrac distanc* test kan de afstandsmetriek laten zien tussen gemeenschappen[7]. De *Bray-Curtis distance* kan aangeven hoeveel gemeenschappelijke soorten er tussen twee populaties zijn[8]. Ook de correlatie tussen soorten kan worden bestudeerd door het uitvoeren van een *Mantel test*[9]. Een *species accumulation curve* kan laten zien hoeveel soorten, en hoeveel individuen er zijn gevonden binnen een bepaald gebied[10]. Al deze toetsen zouden kunnen bijdragen aan het beantwoorden van ons doel. Tijdens het onderzoek zal steeds specifieker duidelijk worden hoe we naar de antwoorden gaan zoeken waarnaar ook duidelijk wordt wat voor testen we zullen uitoefenen op de data.

Onderzoeksopzet

Dataverzameling
De data die wordt gebruik tijdens dit project wordt aangeleverd door Vincent Mercks, een projectleider van ARISE en projectleider van het eerdere onderzoek naar het verkrijgen van deze data. De twee verschillende data zijn aangeleverd rauwe fastq formaat en de andere in een Operational Taxanomic Unit (OTU) tabel. Het OTU-bestand is geen groot formaat, dus kunnen via git worden opgeslagen in de repository. De nieuwe data die wordt aangeleverd is illumina sequenced data. Deze data zijn groter dan het OTU-tabel. Deze data zal worden opgeslagen op een andere locatie. Er zal vanuit Naturalis een soort repository komen ontwikkeld met het iRODS systeem, die grote hoeveelheden data kan opslaan. Er zal een MaaS worden ontwikkeld waar ik in ga werken. Hier zal de data staan en het R-project. Vanaf mijn lokale computer zal toegang zijn tot het werken in de MaaS.
Wetenschappelijke artikelen zullen gebruikt worden om de nodige kennis op te doen voor het onderzoek. Voor het vinden van artikelen wordt PubMed en GoogleScholar gebruikt. Naast wetenschappelijke artikelen is het voorafgaande onderzoek een grote informatiebron, dit is het scriptieverslag van Sophie van Melis.

Data-analyse
De data-analyse zal uitgevoerd worden in een R server omgeving. Met R studio is het mogelijk iCommand en command lines te gebruiken om bijvoorbeeld bij de beschikbare data te komen. Zowel de DADA2 ITS pipeline workflow als alle verwachtte statistische testen zullen uitgevoerd worden in R. Hiervoor zullen meerdere R packages worden gebruikt.

Producten

Tussenproducten
Tussenproducten die verwacht worden van de Hogeschool Leiden:
- Plan van Aanpak
- Posterpresentatie
- Tussenverslag
- Organiseren meeloop dag voor tweedejaars studenten van de opleiding

Op 6 maart 2023 zal er een tussentijdse deadline zijn om je scriptie en voortgang te bespreken met de Hogeschool Leiden. Dit is ook een moment om te kijken of de inzet voldoende is om het onderzoek te kunnen afsluiten. Verder zal er op drie verschillende momenten een tussengesprek zijn met de begeleider van de Hogeschool. Bij het eerste gesprek is de stagebegeleider aanwezig.

Eindproducten
De volgende eindproducten worden verwacht van Naturalis:
1. Het aangeleverde R script om de data met behulp van de DADA2 ITS pipeline workflow om te zetten in een ASV-tabel.
2. Het aangeleverde R script om een rarefaction curve te ontwikkelen.
3. Verschillende scripts om de data te kunnen classificeren.
4. Een opzet voor een manuscript voor een gekozen journal die in gaat op de gevonden resultaten.

Alle eindproducten en gevonden resultaten zullen in manuscript worden opgeschreven. Elk journal heeft andere eisen voor een manuscript. Voor dit onderzoek zal ik een manuscript willen inleveren bij MBC Bioinformatics[11]. Het manuscript zal bij de hogeschool worden ingeleverd als afstudeerscriptie.

In het manuscript komen de volgende hoofdstukken voor:

- '*Title page*'
- *'Abstract'*; een beknopte samenvatting waar in minder dan 350 woorden de achtergrond, resultaten en conclusie worden beschreven
- '*Keywords*'; drie tot tien belangrijkste termen van het manuscript
- '*Background*'; achtergrond van het onderzoek
- '*Results*'; de resultaten van het onderzoek
- '*Discussion*'; een discussie over onderzoeksproblemen die tijdens het onderzoek tegenaan is gelopen
- '*Conclusions*'; de belangrijkste conclusies en de waarde van het onderzoek worden hier besproken
- '*Methods*'; uitgewerkte lijst van de gebruikte technieken en indeling van het onderzoek
- '*List of abbreviations*'; eventuele gebruikte afkortingen worden hier opgeschreven.

Afhankelijk van de bron, zullen de referenties in een behorende stijl worden opgeschreven. De eisen die zijn opgesteld voor de figuren en legenda's[11], zullen worden gevolgd om het manuscript in elkaar te zetten

Projectorganisatie en Begeleiding

Dit project zal worden uitgevoerd door Winny 't Hoen onder leiding van Rutger Vos. De samenwerkingsovereenkomst is als volgt: Er zal drie keer per week een koffie uurtje zijn waarin het mogelijk is om met Rutger, en medestudenten, dingen te kunnen bespreken over je project. Dit is een advies maar geen verplichting. Voor de rest is er contact via de email. Er zal voornamelijk uit huis gewerkt worden, dit is eigen verantwoordelijkheid. Mijn doel is om 2 á 3 keer per week aanwezig te zijn op Naturalis en daar gebruik te maken van de werkplekken. Voor de rest van de week zal ik deels uit huis werken en deels van uit de Universiteitsbibliotheek. Er zal een dagelijkse update zijn op de Git repository. Hier zullen alle updates over het project zijn opgeslagen. Om dit project succesvol te laten verlopen zal ik harde deadlines voor mezelf moeten stellen. Wanneer ik moeite heb met het halen van deze deadlines, zal ik advies vragen bij mijn mede studenten.

Werkplanning

Tijdens dit project zal ik twee belangrijke hertentamens hebben om mijn bachelor te kunnen afronden. Deze hertentamens zijn in de tweede en vierde periode van dit studiejaar. Het tentamen in de tweede periode valt in een drukke periode van mijn stage. Het andere tentamen in periode vier zal tijdens het afronden van mijn stage zijn. Belangrijke inlever momenten staan in het schema als deadline. Per week zal ik een planning voor mezelf maken met details van mijn doelen die week. Dit schema is een ondersteuning bij met maken van deze wekelijkse planning en voor het bijhouden van het verloop van het project.

Periode 1	0	0	1	2	3	4	5	6	7	8	9	10
Datum	22-aug	29-aug	05-sep	12-sep	19-sep	26-sep	03-okt	10-okt	17-okt	24-okt	31-okt	07-nov
Plan van Aanpak schrijven							RV			RV		
DADA2 ITS pipeline workflow toepassen op de nieuwe Illumina dataset							MNL			RV		
Voorbereidingen voor de DADA2 ITS pipeline workflow										RV		
Manuscript format van een journal kiezen										RV		
Vertalen Plan van Aanpak voor manuscript	HS					PvA				RV		
TO DO'S					Gespreksmoment 1 inplannen		Rutger vakantie					

Periode 2	12	13	14	15	16	17	18		19	20	21	22
Datum	14-nov	21-nov	28-nov	05-dec	12-dec	19-dec	26-dec	02-jan	09-jan	16-jan	23-jan	30-jan
Overstap naar de bodemschimmel data							V	RV				
Rarefaction curve opstellen + intepreteren							V	RV				
Classificeren van de data							V	RV				
		HS					V	RV				
BGE herkansing							V	RV				
TO DO'S				Gespreksmoment 2 inplannen								

Periode 3	23	24	25	26	27	28	29	30	31	32	33
Datum	06-feb	13-feb	20-feb	27-feb	06-mrt	13-mrt	20-mrt	27-mrt	03-apr	10-apr	17-apr
Classificeren van de data					Tussenverslag					V	
Werken aan tussen verslag											
Welken aan manuscript											
		HS									
									V		
TO DO'S					Gespreksmoment 3 inplannen						

Periode 4	34	35	36	37	38	39	40	41	42	43	44	45
Datum	24-apr	01-mei	08-mei	15-mei	22-mei	29-mei	05-jun	12-jun	19-jun	26-jun	03-jul	10-jul
Werken aan Manuscript						V						
Voorbereiden op stageverdediging												
Nalezen codes en eventuele opschoning												
Documentatie bijwerken	V			V	HS							
BNGST herkansing		V	Scriptie	V								
TO DO'S						Begin afstudeerzittingen						

Risicoanalyse

Om te voorkomen dat er fouten worden gemaakt waar wij niet op voorbereid zijn wordt er een risicoanalyse opgesteld. In tabel 1 staan de risico's met daarbij de kans op optreden, de impact, de gevolgen en de passende maatregelen.

RISICO	KANS	IMPACT	MAATREGELEN	GEVOLG
1. Tijdsinschatting blijkt te kort	3	2	Accepteren en een nieuwe planning maken	Enkele doelen kunnen misschien niet meer behaald worden
2. De nieuwe data worden laat geleverd	3	2	Planning omdraaien.	Beginnen bij deel 2 van het onderzoek, eindigen met deel 1.
3. Het uitvoeren van deel 1 lukt niet	2	4	Nagaan waarom het niet lukt en aangeven bij begeleiders.	Doorgaan met deel 2.

4.	Concentratievermogen belemmert het onderzoek	3	4	Bespreken met begeleider. Plan aanpassen.	Vertraging van de stage

Risico 1: Tijdens het eerste deel zal de DADA2 ITS pipeline workflow worden gebruikt. Er zijn hele stappenplannen hiervoor te vinden op het internet. Om deze reden gaat mijn verwachting er naar uit dat het niet het grootste deel wordt van het onderzoek. Echter ken ik de data nog niet waarmee ik ga werken, dus zouden er problemen kunnen ontstaan waardoor dit proces langer zou kunnen duren. Het tweede deel van het onderzoek zal naar mijn verwachting langer gaan duren, gezien er veel verschillende kleinere onderzoeken worden uitgevoerd. Ook hiervan kunnen er problemen tussendoor ontstaan die onvoorzien zijn.

Risico 2: Dit is een risico die voornamelijk als gevolg heeft dat het schrijven van mijn scriptie iets uitdagender wordt. Wat fijn is aan zelf de data bewerken met de workflow is dat ik een beter beeld heb van wat er is gebeurd met de data en wat daarvan de gevolgen kunnen zijn. Mocht de nieuwe data laat worden geleverd zal ik vast beginnen bij deel 2.

Risico 3: Deel 1 van het onderzoek is het uitvoeren van de DADA2 ITS pipeline workflow voor illumina data. Er zou natuurlijk altijd iets kunnen zijn met de data waardoor dit proces langer of zelfs te lang duurt. Dan zouden we moeten nagaan of dit nuttig is of dat dit moet worden overgeslagen. Deel 1 is handig voor het onderzoek om te begrijpen wat de stappen zijn geweest die de persoon voor mij met de data heeft gedaan waarmee ik verder ga werken. Echter is het belangrijkste doel om deel 2 van het onderzoek te laten slagen.

Risico 4: Dit risico is voor mij persoonlijk een van mijn grotere angsten wat er tijdens dit project kan gaan gebeuren. Tijdens mijn studie heb ik veel op mijn eigen tempo kunnen indelen, waardoor ik een jaar extra heb gedaan over de eerste drie jaar van mijn bachelor. De reden hiervan is mijn ADHD en dyslexie. Dit is een van de eerste keren dat ik full time mee draai en hoop hetzelfde te kunnen presteren in deze maanden als een gemiddelde andere student. Mocht ik het idee krijgen dat dit niet lukt, zal ik dat op tijd aankaarten bij mijn begeleiders en opzoek gaan naar oplossingen.

Projectgrenzen

Het project is van start gegaan op 5 september 2022 en eindigt rond juni 2023. Tijdens deze periode zal er 36 uur per week worden gewerkt: Maandag t/m donderdag van 09:00 tot 17:00 en vrijdag tot 13:00. Mochten er tussendoor dagen zijn waarbij ik graag een halve dag wil werken, zal ik dat vrijdagmiddag kunnen inhalen. Er zijn duidelijke doelen gesteld voor dit project. Echter, zoals veel onderzoeksprojecten, kunnen er alleen voorspellingen worden gedaan over welke tools je hiervoor nodig hebt. Een grens stellen voor programmeertalen lijkt mij dan ook onhandig. Mocht er een onbekende programmeertaal nodig zijn, dan zal er in overleg worden bepaald het noodzakelijk is voor het project om deze taal te leren. Beide delen van het project zijn belangrijk, echter is het tweede deel onmisbaar bij het afsluiten van het project.

Literatuurlijst

1. https://www.arise-biodiversity.nl
2. https://www.forestresearch.gov.uk/research/metabarcoding/
3. Sophie van Malis, “DNA metabarcoding analysis of fungal diversity in soil of three vegetations in the Netherlands”, 2022
4. Gavito et al., “Local-scale spatial diversity patterns of ecotomycorrhizal fungal commities in a subtropical pine-oak forest”, Fungal Ecology, 2019, 42
5. https://benjjneb.github.io/dada2/ITS_workflow.html
6. https://cran.r-project.org/web/packages/Rarefy/vignettes/Rarefy_basics.html
7. https://joey711.github.io/phyloseq-demo/unifrac.html
8. **Stephanie Glen**. "Bray Curtis Dissimilarity" From **StatisticsHowTo.com**: Elementary Statistics for the rest of us! https://www.statisticshowto.com/bray-curtis-dissimilarity/
9. https://jkzorz.github.io/2019/07/08/mantel-test.html
10. https://terrestrialecosystems.com/species-accumulation-curves/
11. https://bmcbioinformatics.biomedcentral.com