15-5-2018

Evander van Wolfswinkel, Rick Schoenmaker en Valerie Verhalle

HAN, Bio-inforamtica, leeraar 1

**Onderzoeksplan**

Micro-organismen in compost voor champignonteelt

[](https://www.google.nl/url?sa=i&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwjY9Jj61ofbAhWKJVAKHcttCAQQjRx6BAgBEAU&url=https://www.belg.be/44536/nieuws/champignonteelt-krijgt-400-000-euro/&psig=AOvVaw0v3pLgGqtqUueiNggMv1mx&ust=1526472392193579)

***Inhoud***

[Inleiding 2](#_Toc514662417)

[Onderzoeksvragen 3](#_Toc514662418)

[Beschrijving onderzoeksplan 4](#_Toc514662419)

[Stappenplan 6](#_Toc514662420)

[Bronnen 8](#_Toc514662421)

[Bijlagen 9](#_Toc514662422)

Inleiding

De champignon (Agaricus bisporus) is een schimmel uit de familie Agaricaceae. Bij de productie van champignons wordt speciale compost gebruikt die bestaat uit een mengsel van onder andere paardenmest, kippenmest, stro en gips. Het proces waarbij de compost geproduceerd wordt bestaat uit 3 fases. In de eerste fase worden de verschillende grondstoffen gemengd en bevochtigd. Waarna in fase 2 dit mengsel gepasteuriseerd wordt om ongewenste micro-organismen te doden. Vervolgens wordt in fase 3 de compost geënt met het mycelium van de champignon (Agaricus bisporus) en doorgroeit het mycelium de compost in ruim 2 weken.

Een groot probleem binnen de champignonteelt is echter dat de oogst van champignons sterk kan variëren. Alle micro-organismen in de compost vormen samen een microflora. De microflora kan een positieve of negatieve rol spelen op de micro-organismen in de compost. Er ontbreken op dit moment methoden om aanwezige micro-organismen in de compost te detecteren, karakteriseren en te kwantificeren, waardoor er momenteel weinig inzicht in de aanwezigheid en samenstelling van microflora in de verschillende fases van het proces en de daarbij behorende invloed die de microflora heeft op teeltresultaten.

Metagenomics (National Institutes of Health, 2007) is hiervoor de oplossing. Metagenomics zijn meerdere verschillende onderzoekstechnieken om zo de microflora uit een omgevingssample vast te kunnen stellen. Dan wordt vastgesteld uit welke micro-organismen deze microflora bestaat. Dit wordt gedaan middels een programma, dat in staat is om meerdere verkregen sequenties te kunnen analyseren. Aan het eind dit onderzoek is er een overzicht van welke micro-organismen zich in het sample van de compost bevonden maar ook interessante genen die ze bevatten en eventueel belangrijke eiwitten die geproduceerd worden. Met deze resultaten kunnen dan bijvoorbeeld biologen aan de slag om een methode te ontwikkelen die zorgt voor meer constante teeltresultaten in de champignonteelt.

Onderzoeksvragen

Tijdens dit onderzoek willen we antwoord verkrijgen op de volgende hoofd- en deelvragen:

*Hoofdvragen*:

* Welke micro-organismen kun je identificeren aan de hand van de dataset?
* Welke eiwitten kun je identificeren aan de hand van de dataset?

*Deelvragen*:

* Door welke genen worden de eiwitten geproduceerd?
* Wat zijn de functies van de gevonden eiwitten?
* Welke domeinen bevatten de eiwitten?
* Wat is de locatie van de eiwitten in de cel?
* Wat is de topologie van de eiwitten?
* Waar komen de micro-organismen voor?
* Welke eigenschappen hebben deze micro-organismen?

Beschrijving onderzoeksplan

Voor het onderzoek gebruiken we de sequentie data die is aangeleverd door het HAN BioCentre. Deze data is verkregen doormiddel van een metagenomics onderzoek naar de verschillende micro-organismen in de compost die gebruikt wordt om champignons te kweken.

Voor het analyseren van de sequentie data gebruiken we de tool Blast met als algoritme Blastx (Ian Korf, 2003-07-29) van de NCBI database. De keuze voor Blastx komt omdat dit naar verwachting de meest nauwkeurige resultaten geeft binnen de beperkte tijd die het onderzoek duurt. Mochten de resultaten van Blastx der mate onnauwkeurig zijn, dan stappen we over op tBlastx dit neemt meer tijd in beslag maar is nauwkeuriger.

Om de nauwkeurigheid van kwaliteit van de alignments te verifiëren gebruiken we een scorematrix. Hierbij gaan we er vanuit dat de meeste sequenties verwant zijn, maar we gaan er ook vanuit dat er micro-organismen tussen zitten die geen verwantschap delen. BLOSUM62 zal gebruikt worden omdat dit de beste balans is tussen de veruit elkaar gelegen sequenties en de dichter bij elkaar gelegen sequenties. Bij Blastx gebruiken we als database de non-redundant GenBank, eukaryoten worden uitgesloten van de resultaten. Dit omdat dit onderzoek zich alleen richt op de micro-organismen en de mogelijk interessante eiwitten die zij produceren en daarmee van invloed kunnen zijn op de champignongroei. Voor elke gap in de sequentie worden er 10 minpunten gerekend voor het openen van een gap en 1 minpunt voor elke positie in de gap. Bacteriën muteren veel door hun hoge reproductiesnelheid waardoor er veel varianten ontstaan. Om deze reden wordt er iets soepeler geteld in het openen van een gap. De ‘word size’ zal ingesteld woorden op 6, ‘low complexity filter’ staat aan omdat repeats niet mee genomen worden in de Blast resultaten. Bacteriën bevatten geen repeats in hun genoom dus hier wordt geen rekening mee gehouden. Het opslaan van de gevonden matches doen we met een “cut-off” op een E-value van 1\*10-10.

Alle gevonden matches met een lagere E-value worden als te onnauwkeurig beschouwd en daarom niet meegenomen in het verdere onderzoek. Andere “cut-off” criteria zijn een minimale identity van 75% (bij eiwitten 25%), een positive percentage hoger dan 70% en een coverage hoger dan 80%. Identity geeft aan hoe goed de gevonden hit is, coverage is een indicator voor hoe goed de alignment is en de positive percent geeft aan hoe goed aminozuren op elkaar lijken qua vorm, lading en interacties. Alle 3 deze criteria zijn van belang voor een goede match.

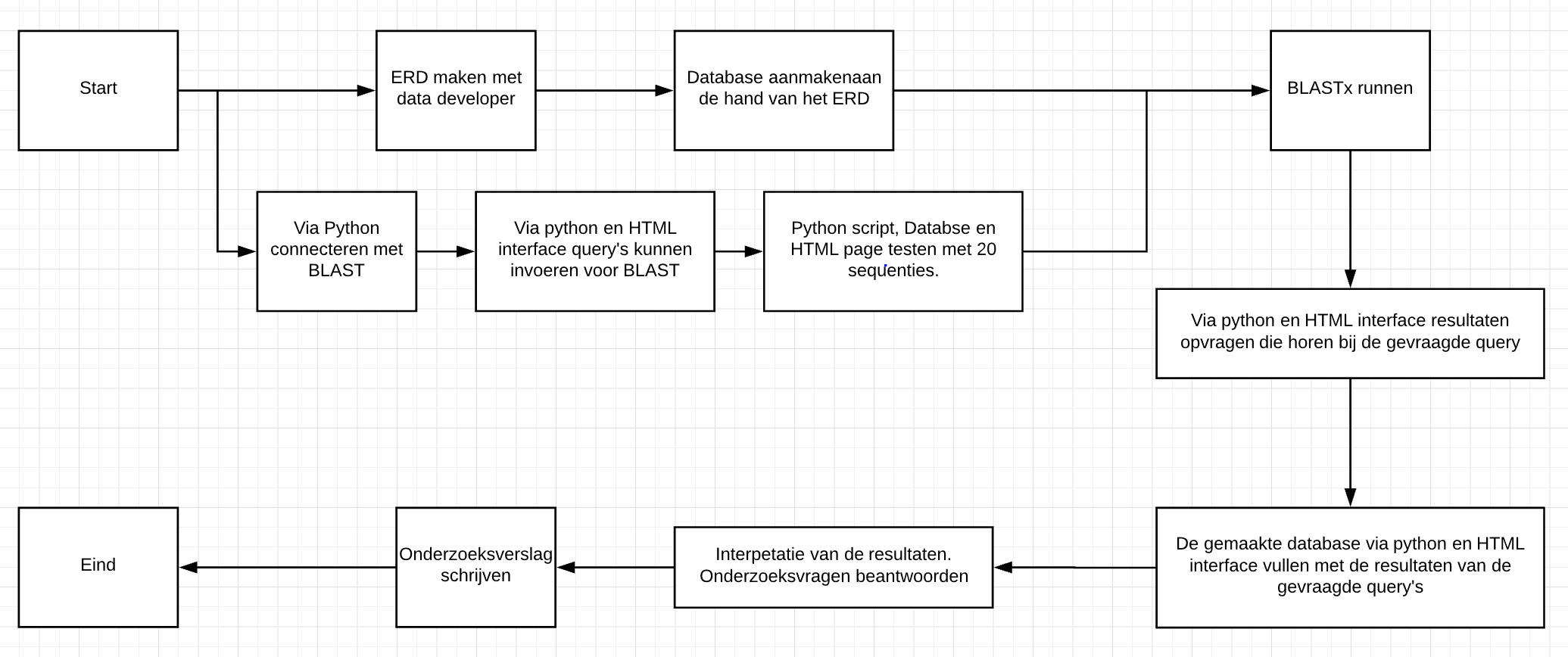
Alleen de eerste 10 gevonden matches die aan deze criteria voldoen worden opgeslagen in een Oracle database met behulp van datamodeler (Kolbe, April 2015) voor het maken van een ERD en SQL developer (Narayanan, Januari 2016) voor de opbouw van de database. De opzet van onze database is als coneptueel ERD weergegeven in de bijlagen (Bijlage1). Doormiddel van een programma geschreven in de programmeertaal Python (Rossum, 1995) en de modules BioPython (Peter J. A. Cock, march 2009) en Matplotlib (Willems, 2017) en de webapplicatie Flask (Grinberg, 2014) worden de resultaten verwerkt. Deze data is vervolgens uit te lezen via een website die weer met behulp van Python verbonden is met de database. De opzet van deze website word geschreven in HTML.

De website laat de gebruiker de verschillende micro-organismen opvragen, de Blast resultaten en eventuele match resultaten die voldoen aan de gevraagde query. Om de applicatie te testen worden er 10 forward en 10 reverse “reads” geautomatiseerd geblast en vervolgens via een ander script automatisch ingeladen in de Oracle database. Doordat deze “reads” bekend zijn kan er aan de hand van de resultaten in de database de conclusie getrokken worden dat het programma werkt.

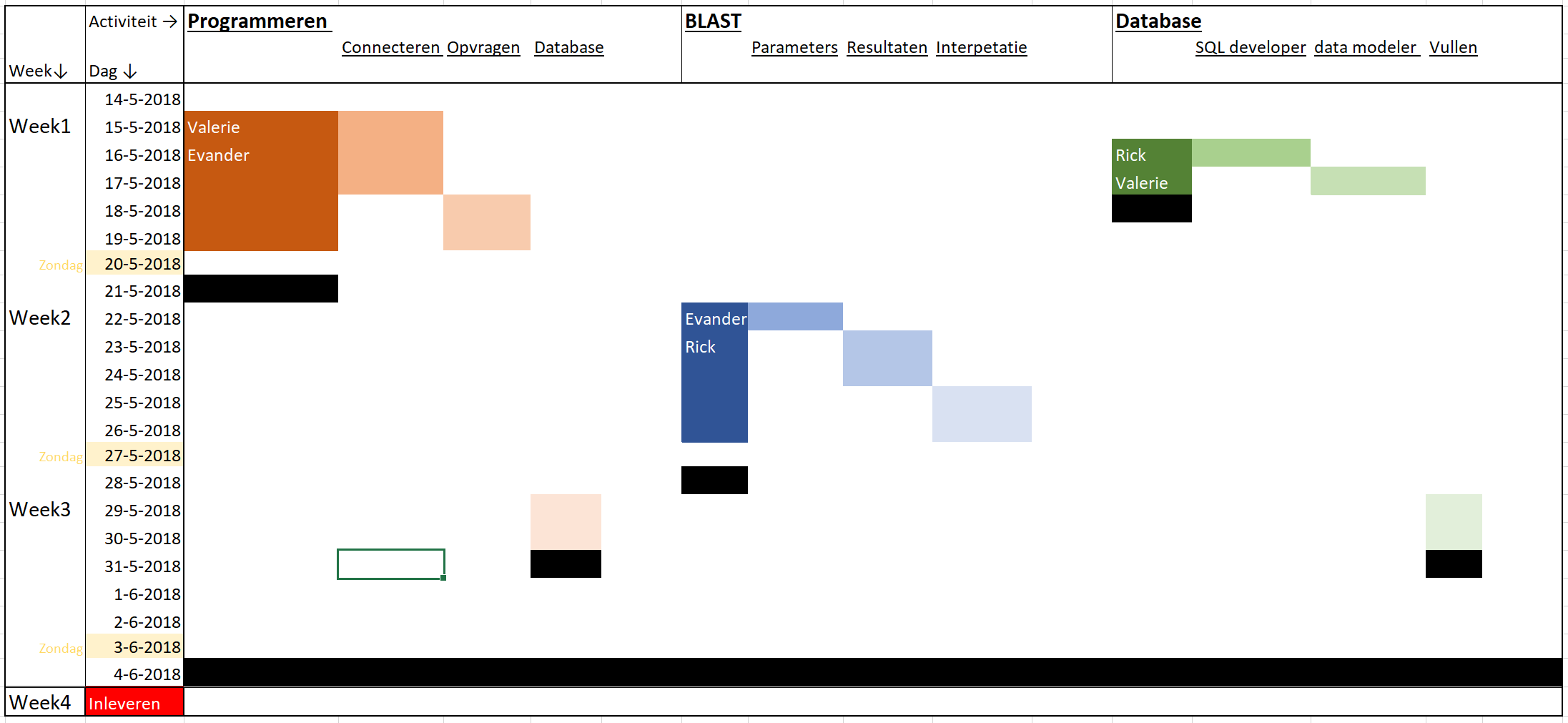
Stappenplan

In figuur 1 (flowchart) en 2 (werkschema) zijn overzichtelijk weergegeven wie er welke handelingen uitvoeren maar ook wanneer deze handelingen uitgevoerd zullen worden.

1Flowchart, overzicht van wanneer welke handelingen gedaan zullen worden.

Figuur 1

2Werkschema, overzicht van wie welke handelingen zal gaan uitvoeren, uitgezet tegen het aantal dagen.

Figuur 2

**Programmeren**:

* Connecteren
  + Via python connecteren met BLAST.
* Opvragen
  + Via python en HTML interface query’s kunnen invoeren voor BLAST.
* Database
  + De gemaakte database via python en HTML interface kunnen vullen met de resultaten van de gevraagde query’s.

**BLAST**:

* Parameters
  + Zorgen dat de parameters vaststaan en kloppen.
* Resultaten
  + Blasten en de resultaten opslaan.
* Interpretatie
  + Interpretatie van de resultaten, hoofd en sub-vragen beantwoorden.

**Database**

* Data-developer
  + ERD maken.
* SQL-developer
  + Database aanmaken aan de hand van het ERD.
* Invullen van de database
  + De database vullen aan de hand van de via BLAST verkregen resultaten

Bronnen

Altschul, S. G.-4. (sd). *tBlastx NCBI*. Opgehaald van NCBI: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=tblastx&BLAST\_PROGRAMS=tblastx&PAGE\_TYPE=BlastSearch&SHOW\_DEFAULTS=on&LINK\_LOC=blasthome

Grinberg, M. (2014). *Flask Web Development: Developing Web Applications with Python.* O'Reilly Media, Inc.

Ian Korf, M. Y. (2003-07-29). *BLAST: An Essential Guide to the Basic Local Alignment Search Tool, first edition.* O'Reilly Media. Opgehaald van NCBI: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastx&PAGE\_TYPE=BlastSearch&LINK\_LOC=blasthome

Kolbe, H. H. (April 2015). *Oracle SQL Developer Data Modeler for Databse Design Mastery, 1st edition.* Europa: McGraw-Hill Education group.

Narayanan, A. (Januari 2016). *Oracle SQL Developer, 1st edition.* Packt Publishing.

National Institutes of Health. (2007). *The New Sciece of Metagenomics: Revealing the Secrets of Our Microbial Planet.* Washington (DC): National Academies Press (US). Opgehaald van NCBI: https://www.ncbi.nlm.nih.gov/books/NBK54011/

Peter J. A. Cock, T. A. (march 2009). *Biopython: freely available Python tools for computational molecular biology and bioinformatics, 1st edition.* ISCB (International Society for Computational Biology).

Rossum, G. (1995, May). *Python tutorial, 1st edition.* Amsterdam, The Netherlands: CWI (Centre for Mathematics and Computer Science).

Willems, K. (2017, February 22). *Matplotlib Turtorial: Python Plotting*. Opgehaald van DataCamp: https://www.datacamp.com/community/tutorials/matplotlib-tutorial-python

Bijlagen

Bijlagen1

De database zal aan BLASTe hand van het volgende conceptueel ERD worden gemaakt:

