**Technische Documentatie**

Bart Jolink, Wouter Gaykema en Maite van den Noort

*13-06-2019*

Inhoud

[1. Inleiding 2](#_Toc11424856)

[2. Werking van de applicatie 3](#_Toc11424857)

[3. Database 8](#_Toc11424858)

[4. Software architectuur 9](#_Toc11424859)

[5. Testscripts en resultaten 10](#_Toc11424860)

[6. Literatuurlijst 12](#_Toc11424861)

[Bijlagen 13](#_Toc11424862)

[Bijlage I Overige scripts 13](#_Toc11424863)

# Inleiding

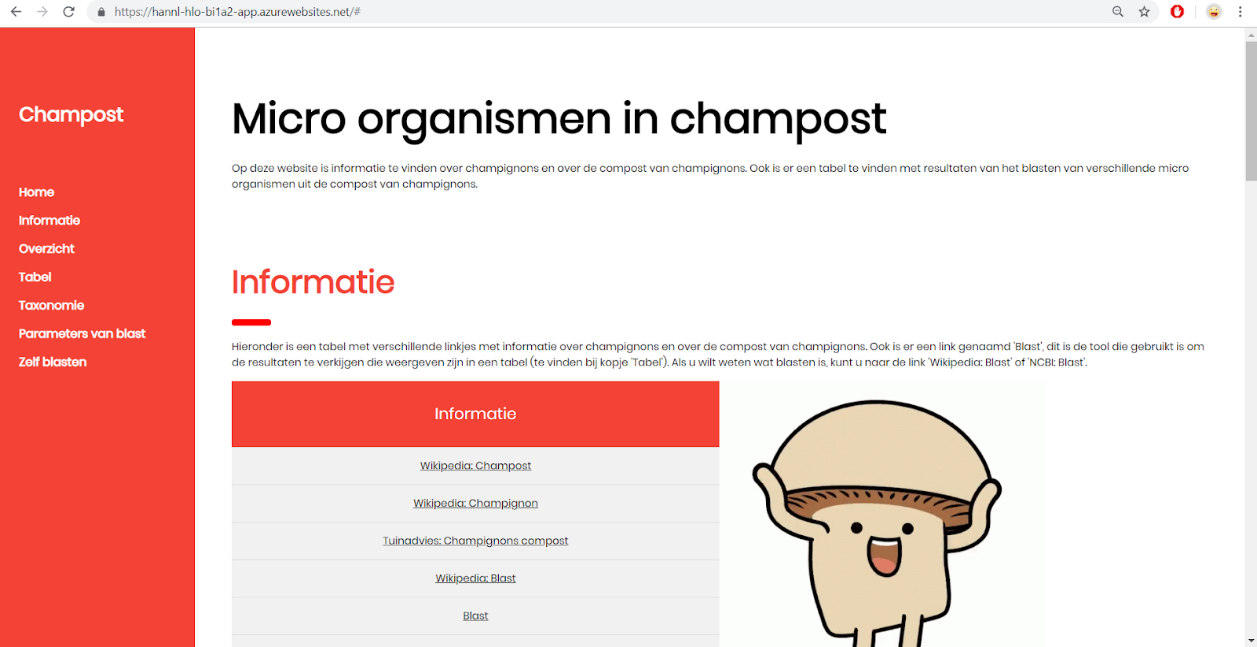
Om de effecten van de samenstelling van de microflora in compost op de uiteindelijke opbrengst aan champignons te kunnen verbeteren is er aan het HAN Biocentre gevraagd om deze samenstelling te bestuderen. Hiervoor is er een metagenomics onderzoek uitgevoerd, waarna DNA gesequenced is middels paired-end illumina sequencing [Ravi et al. 2018]. De reads zijn opgeslagen in FastQ format [Cock et al. 2010], bestaande uit een header, sequentie en een score (in ASCII-code) die staat voor de kwaliteit van de read. Deze reads zijn vervolgens automatisch aligned via Basic Local Alignment Search Tool (BLAST) [Altschul et al. 1990]. Resultaten van deze BLAST zijn vervolgens ook automatisch opgeslagen in een database. De database is te raadplegen via een ontwikkelde applicatie (<https://hannl-hlo-bi1a2-app.azurewebsites.net/>), waarbij alle relevante informatie op een manier getoond wordt zodat deze behapbaar is voor biologen. In deze documentatie zal het script van de applicatie, en de werking daarvan, beschreven worden.

Scripts voor het automatisch blasten, het vullen van de database en het opvragen van de taxonomie zijn vooraf gebruikt om de data te genereren welke de applicatie aanroept. Informatie hierover is terug te vinden in Bijlage I.

# Werking van de applicatie

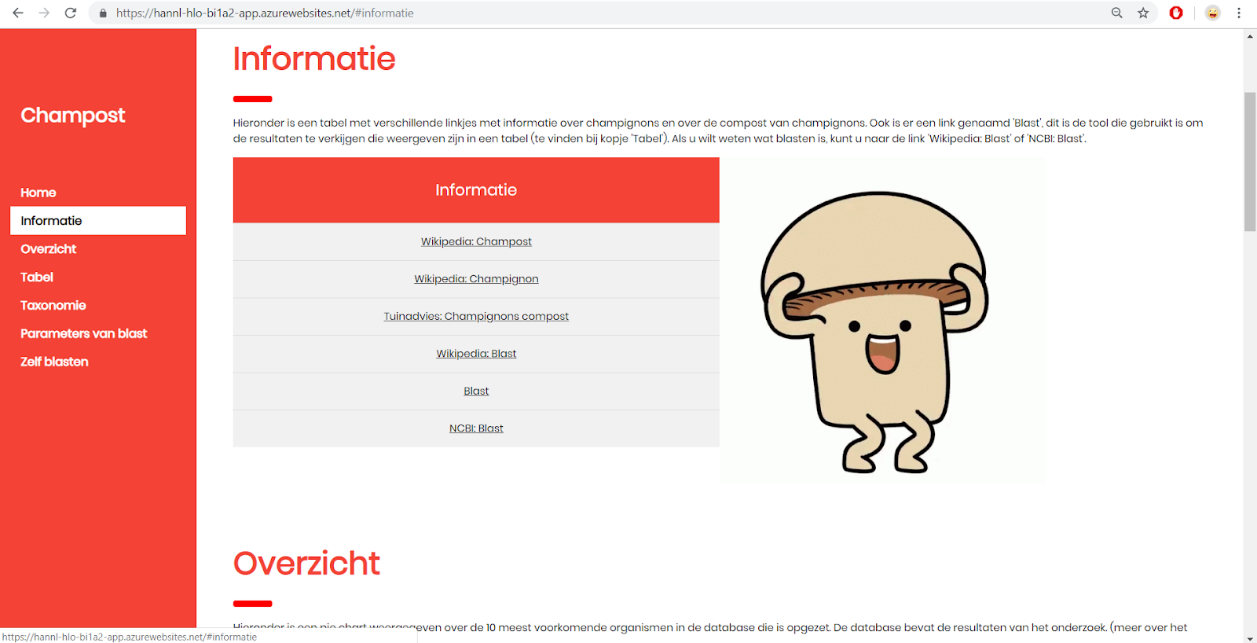
De doelstelling van deze applicatie is het overzichtelijk weergeven van de resultaten uit het onderzoek naar de microflora in compost van champignons. De eerste stap voor het raadplegen van de applicatie is het volgen van de volgende url: <https://hannl-hlo-bi1a2-app.azurewebsites.net/>

Na het volgen van de URL wordt een pagina als op figuur 1 getoond. Op deze pagina kan gescrold worden om naar het volgende kopje (bijv. Informatie, Tabel, etc.) te gaan, maar er kan ook aan de linkerkant op geklikt worden om naar het desbetreffende kopje te navigeren. Deze methode van navigeren, waarbij wordt geklikt, wordt aangeraden en zal dan ook verder in deze documentatie gebruikt worden. Bij het klikken op ‘Home’ krijgt u tevens onderstaande pagina te zien.



Figuur 1 Weergave Homopage. Eigen werk auteur.

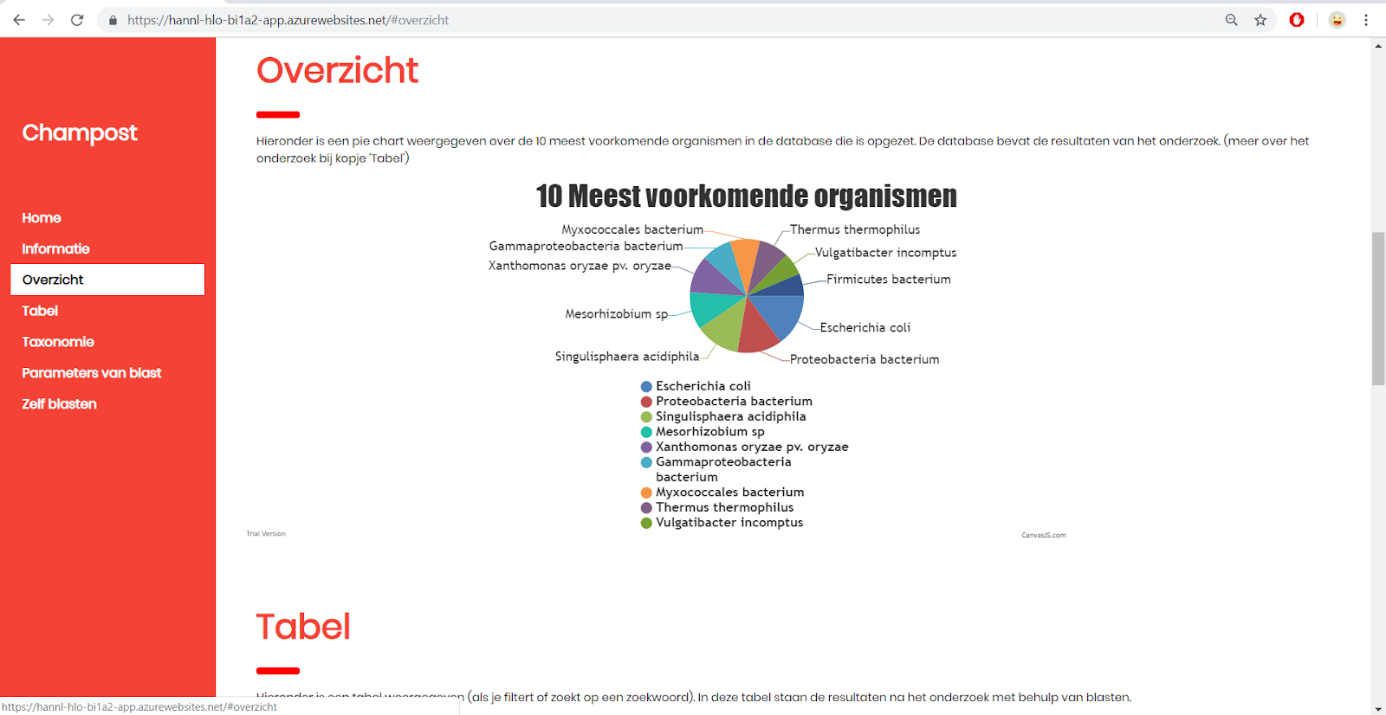
Bij het klikken op ‘Informatie’ zal de pagina zich naar boven of beneden verschuiven (afhankelijk van vorige positie) zodat de pagina wordt weergegeven als in figuur 2. Bij Informatie staan enkele handige links naar relevante informatie omtrent het onderzoeksgebied.



Figuur 2 Weergave kopje 'Informatie'. Eigen werk auteur.

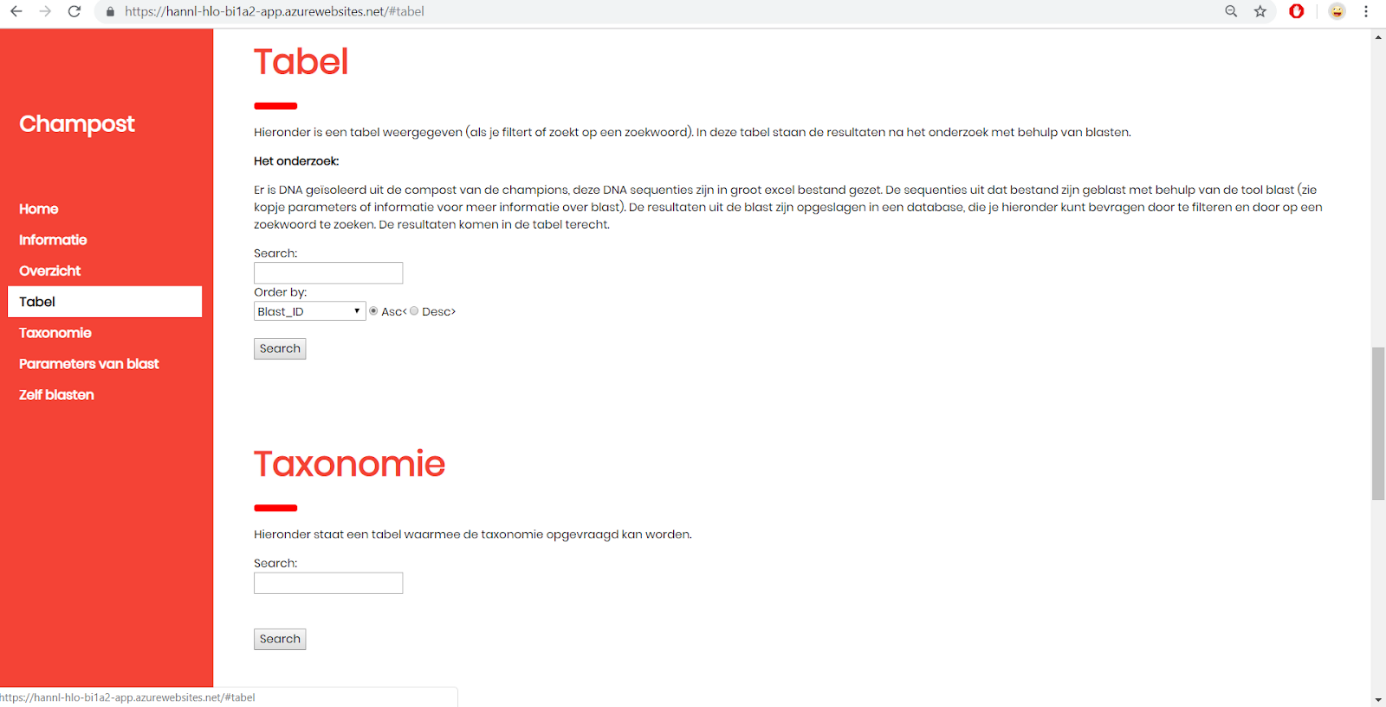
Overige kopjes kennen een soortgelijke navigatie, waarbij de pagina zich naar boven of beneden verschuift, zodat het desbetreffende kopje boven aan de weergegeven pagina staat.

Bij het klikken op ‘Overzicht’ krijgt u figuur 3 te zien. In dit overzicht worden de 10 meest voorkomende organismen weergegeven in een pie chart.



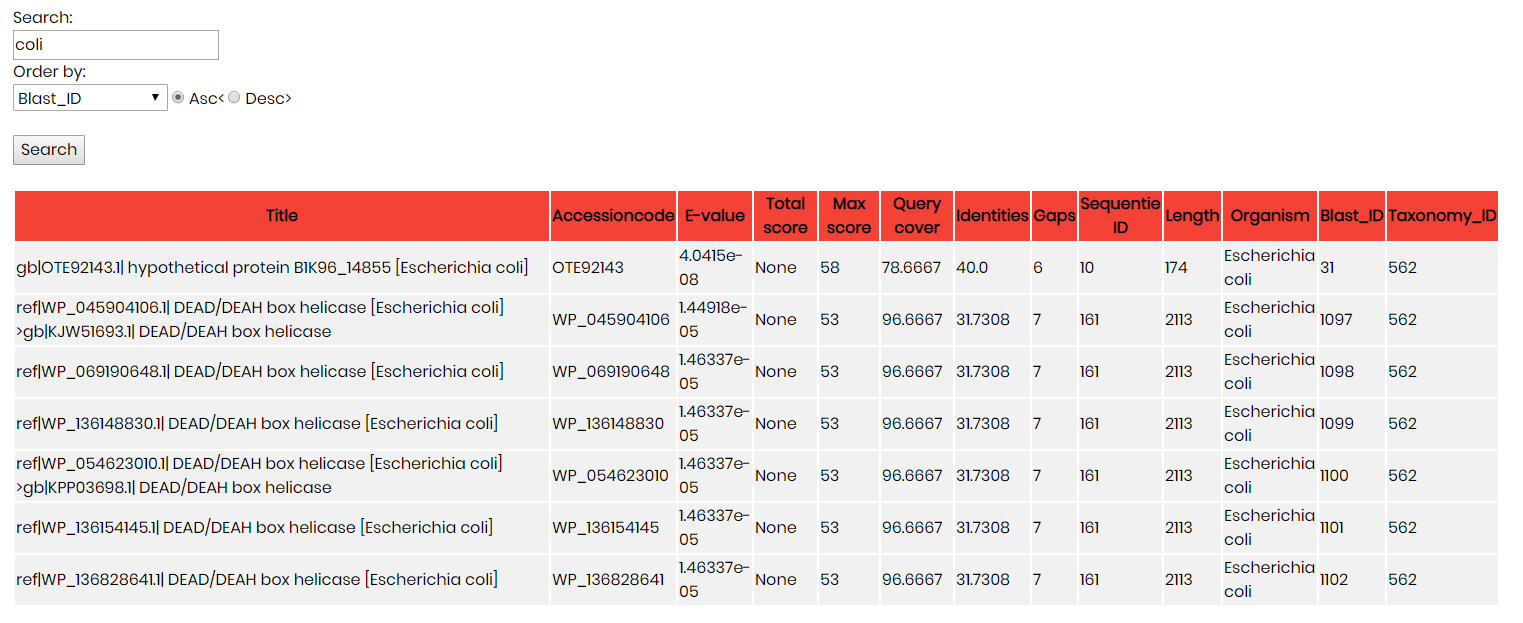
Figuur 3 Weergave kopje 'Overzicht'. Eigen werk auteur.

Bij het klikken op ‘Tabel’ krijgt u figuur 4 te zien. Deze tabel kan gebruikt worden om de informatie uit de BLAST resultaten raad te plegen. In de zoekbalk kunt u uw zoekterm invullen, deze zoekt op titel, organisme en accessiecode. In het ‘Order By’ dropdown-menu kunt u selecteren op welke wijze de gevonden resultaten gesorteerd worden. Dit kan oplopend (Asc<) of aflopend (Desc>). Wanneer vervolgens de gewenste gegevens zijn ingevuld, kunt u op ‘Search’ klikken om de resultaten op te vragen. Standaard zullen de resultaten oplopend gesorteerd worden op Blast\_ID. Bij het leeglaten van de zoekbalk zullen alle resultaten getoond worden.



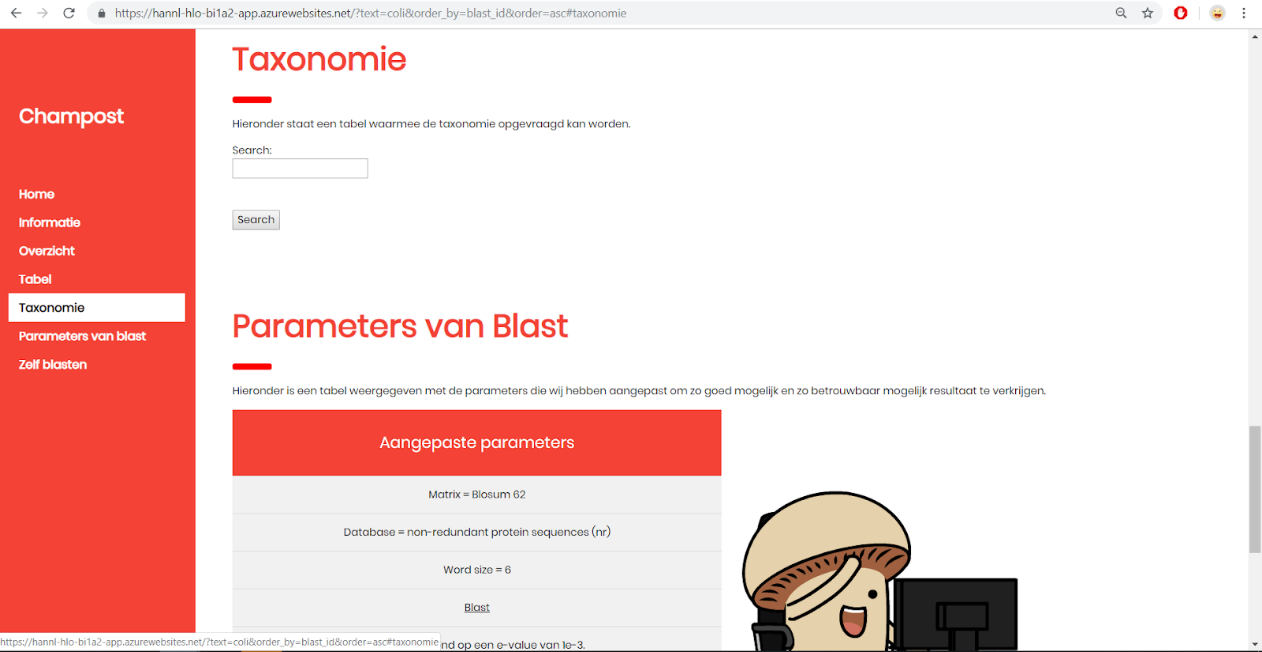
Figuur 4 Weergave kopje 'Tabel'. Eigen werk auteur.

In figuur 5 zijn de resultaten te zien van een zoekopdracht waarbij gezocht werd op ‘coli’. Zoals te zien is, is deze oplopend gesorteerd op Blast\_ID. In deze tabel zijn meerdere kolommen met informatie te zien; Title, Accessioncode, E-Value, Total score, Max score, Query cover, Identities, Gaps, Sequentie ID, Length, Organism, Blast\_ID en Taxonomy\_ID.



Figuur 5 Weergave resultaten tabel bij zoekopdracht 'coli'. Eigen werk auteur.

Bij het klikken op ‘Taxonomie’ krijgt u figuur 6 te zien. Deze taxonomie tabel kan gebruikt worden om de taxonomische rangschikking van een organisme te raadplegen. In de zoekbalk kan een zoekopdracht ingevuld worden. Indien er maar één resultaat voldoet aan de zoekopdracht, dan zal de taxonomische rangschikking vanaf die rang omhoog gegeven worden. Zoals te zien is in figuur 7 zal bij het ingeven van een species naam deze de rangschikking vanaf species weergeven, bij het ingeven van een orde naam zal deze de rangschikking vanaf orde weergeven. Indien er meerdere resultaten voldoen aan de zoekopdracht, dan zullen alle hits weergegeven worden in een tabel. Hierbij wordt de gebruiker geacht de zoekopdracht te specificeren zodat er maar één naam gekozen wordt. Een voorbeeld hiervan is te zien in figuur 8.



Figuur 6 Weergave kopje 'Taxonomie'. Eigen werk auteur.

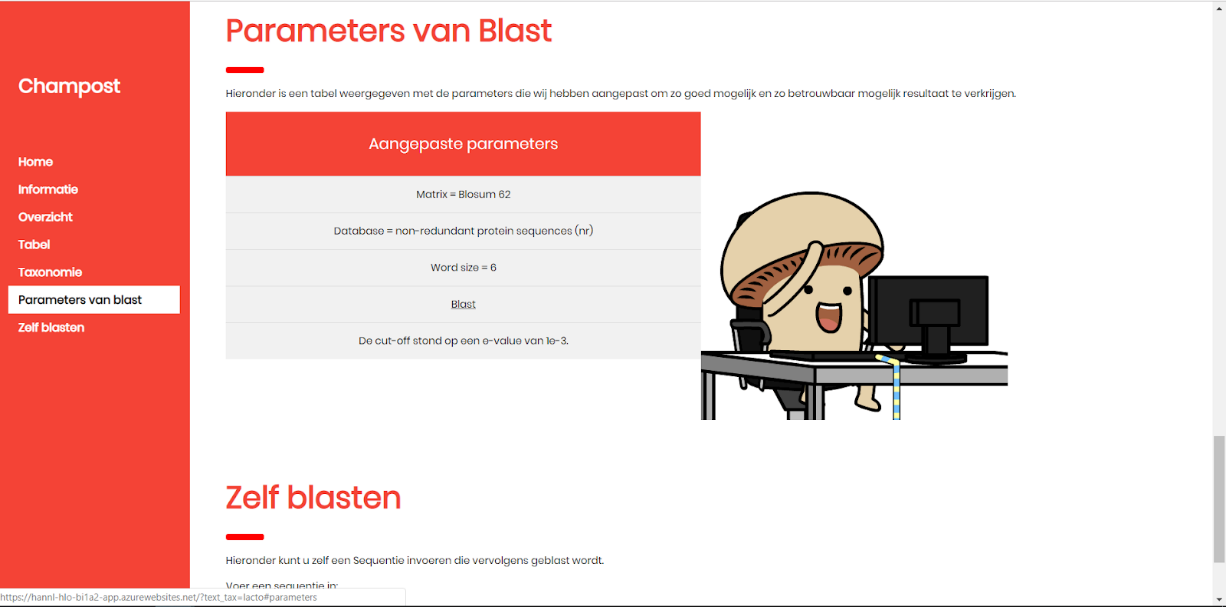


Figuur 7 Weergave resultaat zoekopdracht voor een species (rechts) en een orde (links). Eigen werk auteur.



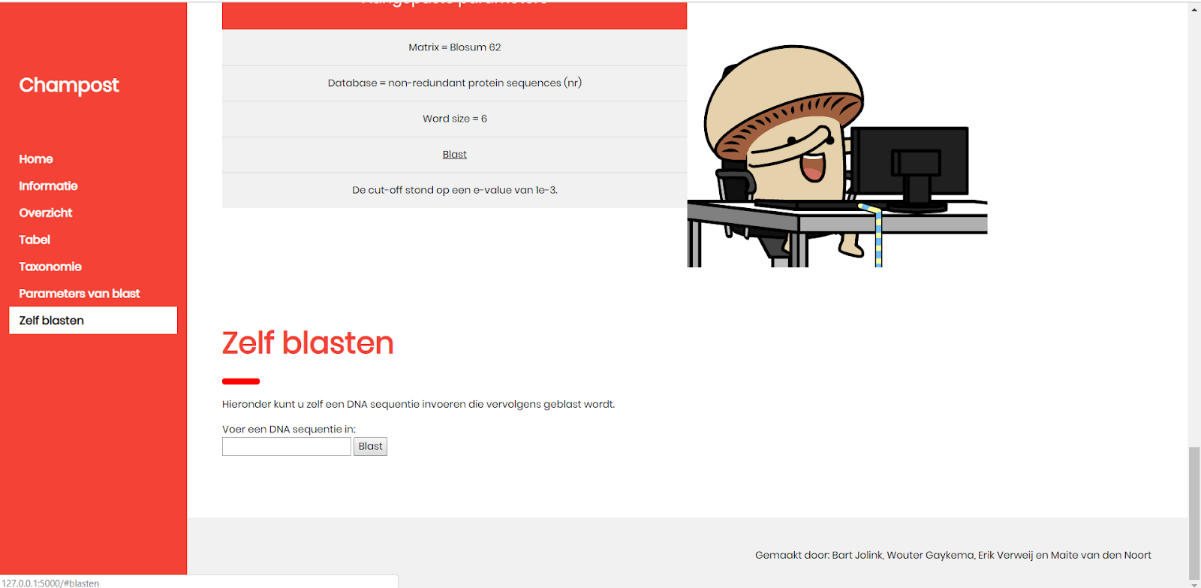
Figuur 8 Weergave zoekopdracht met meerdere resultaten. Eigen werk auteur.

Bij het klikken op ‘Parameters van blast’ krijgt u figuur 9 te zien. Hierbij wordt enkele informatie verschaft over de gebruikte parameters voor de blast.

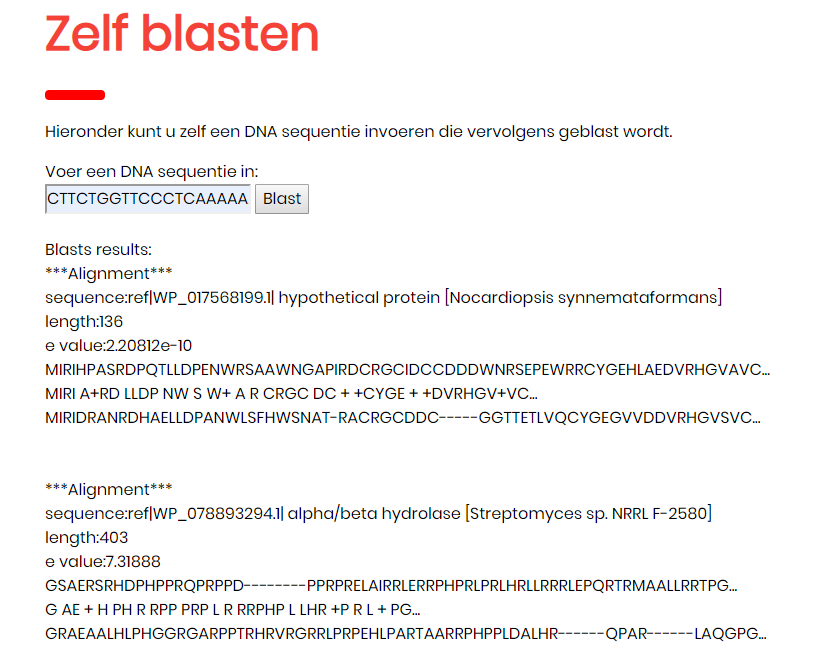


Figuur 9 Weergave kopje 'Paramaters van blast'. Eigen werk auteur.

De laatste functionaliteit is raad te plegen door te klikken op ‘Zelf blasten’. De pagina wordt weergegeven zoals in figuur 10. Hierbij kan een DNA sequentie ingevoerd worden, welke vervolgens doormiddel van BLASTx geblast wordt. De resultaten (max. vijf) worden vervolgens weergegeven zoals in figuur 11. ‘sequence’ staat voor de aligned sequentie, ‘length’ voor de lengte van de alignment en ‘e-value’ voor e-value van de alignment. Verder wordt een deel van de alignment zelf weergegeven, waarbij de bovenste regel de ingevoerde sequentie is (vertaald naar aminozuren), de middelste regel de similiarties en de onderste regel de gematchte sequentie. Deze resultaten zullen niet toegevoegd worden aan de huidige database, om vervuiling van de database te voorkomen.

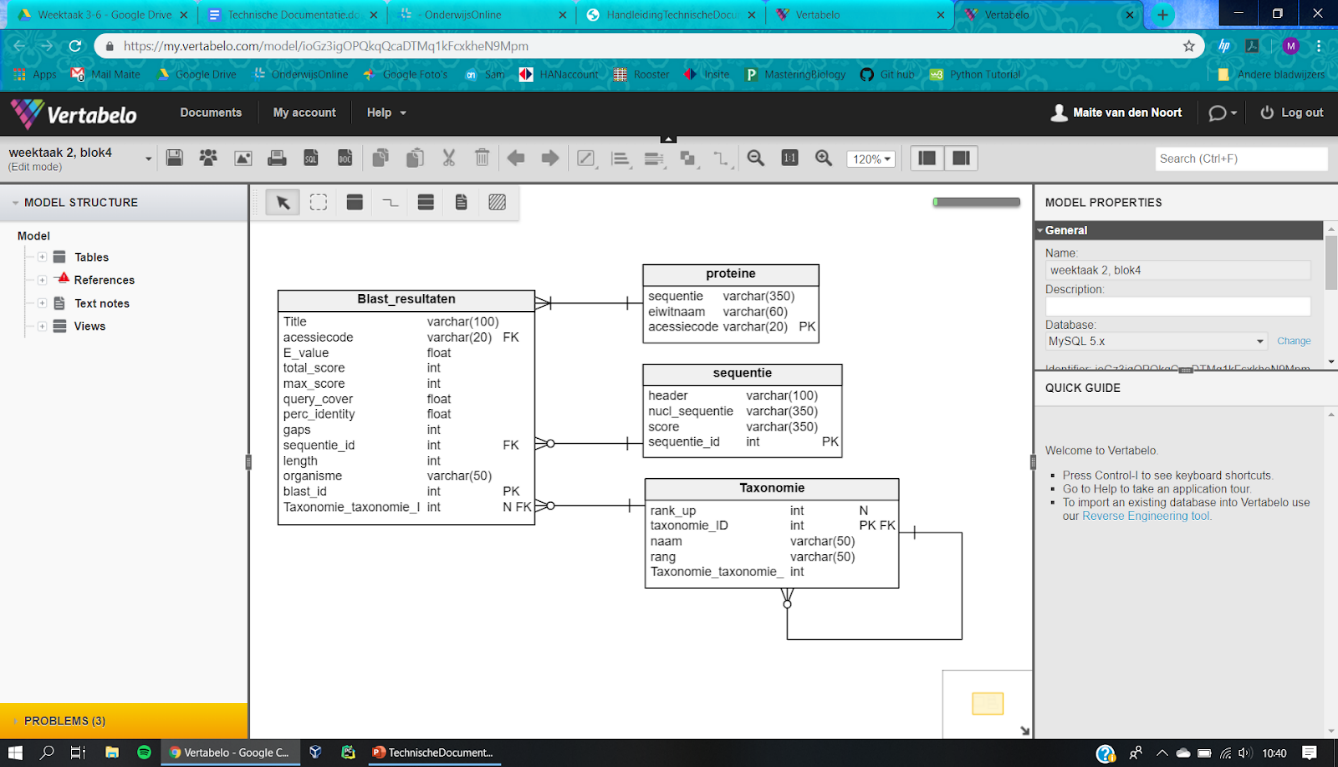


Figuur 10 Weergave kopje 'Zelf blasten'. Eigen werk auteur.



Figuur 11 Weergave resultaten van blast. ‘sequence’ staat voor de aligned sequentie, ‘length’ voor de lengte van de alignment en ‘e-value’ voor e-value van de alignment. Verder wordt een deel van de alignment zelf weergegeven.

# Database



Figuur 12 Technisch ERD MySQL database

Voor het opslaan van de data is gebruik gemaakt van een MySQL [Widenius et al, 2002] database. Hierbij zijn er vier tabellen gemaakt; Blast\_resultaten, proteïne, sequentie en Taxonomie. Figuur 12 laat het technisch Entitiy Relationship Diagram (ERD) van deze database zien.

Er is gekozen voor een één-op-veel relatie tussen de tabel proteïne en Blast\_resultaten omdat een accessiecode meerdere blast resultaten kan hebben en elke blast resultaat is gelinkt aan één accessiecode. Ook is er gekozen voor een één-op-veel tussen de tabel sequentie en Blast\_resultaten omdat elk sequentie\_id meerdere resultaten kan hebben. Het zou ook kunnen dat er helemaal geen resultaten terug komen. Een blast resultaat is altijd gelinkt aan één sequentie\_id. Daarnaast is er gekozen voor een één-op-veel relatie tussen Taxonomie en Blast\_resultaten omdat één taxonomie\_ID meerdere blast resultaten kan hebben, het kan ook zijn dat er geen resultaten uit de blast komen. De Blast\_resultaten hebben elk maar één taxonomie\_ID. Ook heeft de taxonomie een één-op-veel relatie met zichzelf, dit is uit te leggen met een voorbeeld: een orde heeft meerdere families, maar een familie heeft maar één orde. Hetzelfde geldt voor geslacht en familie; een familie heeft meerdere geslachten maar een geslacht behoort maar tot één familie. En zo is het met de hele taxonomische rang en is er dus gekozen voor een één-op-veel relatie met zichzelf.

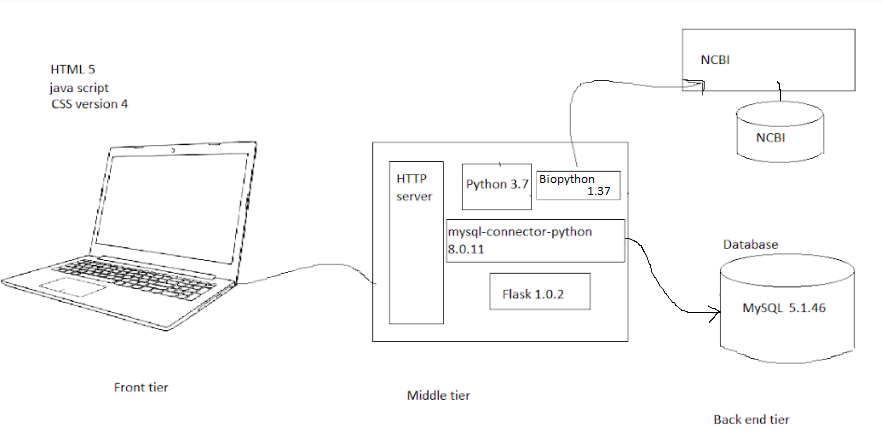
De tabel Blast\_resulaten bevat alle resultaten van een geblaste sequentie en is zodanig opgezet dat er flexibel uitgebreid kan worden. De tabel proteïne bevat alle nodige informatie over de bijhorende proteïne van de blast resultaten. De tabel sequentie bevat alle nodige informatie over de geblaste sequentie zelf, de blast resultaten zelf staan niet in deze tabel. De tabel Taxonomie bevat alle nodige informatie over de taxonomische rangschikking van de gevonden organismen.

# Software architectuur

De gebruiker maakt, door middel van een HyperText Transfer Protocol Secure (HTTPS) server, een connectie met de website (<https://hannl-hlo-bi1a2-app.azurewebsites.net/>), welke gehost wordt op Microsoft Azure [Copeland, et al. 2015]. Vanuit deze site kunnen meerdere functies worden aangeroepen.

De applicatie (app.py) wordt opgevraagd vanuit een Python 3.7 script waarbij gebruik is gemaakt van Flask 1.0.2 [Lord et al, zd]. Flask kan HTML code retourneren terwijl je ook de functionaliteiten van Python behoudt. Het Python script retourneert een van meerdere ‘versies’ van HTML-5 code, op basis van de gevraagde functie via de website. Ook bevat de HTML code een deel Javascript, wat gebruikt is voor de pie chart, en verwijzingen naar CSS 4 files (via linkjes). De code (app.py) is raad te plegen via <https://github.com/BartJolink/Azure-Bi1a2/blob/master/app.py>.

De Python code maakt gebruik van de mysql-connector-python 8.0.11 [Oracle. 2019] om te connecteren aan de MySQL 5.1.46 database waarin gegevens omtrent de blast resultaten staan opgeslagen. Ook kan de Python code, via de BioPython 1.37 module [Chapman en Chang, 2000], de National Center for Biotechnology Information (NCBI) database [O’Leary et al. 2016] raadplegen. BioPython bevat een functie om vanuit Python script te blasten, welke gebruikt wordt voor het ‘zelf blasten’ op de website.



Figuur 13 Software architectuur van applicatie. Via de front tier wordt de applicatie (middle tier) aangeroepen, welke informatie haalt uit de back-end tier.

# Testscripts en resultaten

De applicatie is raad te plegen via: <https://hannl-hlo-bi1a2-app.azurewebsites.net/>.

Op de site staat een toolbar met zeven verschillende kopjes. Waar in deze technische documentatie  op wordt gefocust is ‘Tabel’, ‘Taxonomie’ en ‘Zelf blasten’; de andere kopjes bevatten slechts wat extra informatie.

Alle functies zijn getest op 13 juni 2019.

Tabel 1 Werking van 'Tabel'

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Stap** | **Instructie** | **Verwacht resultaat** | **Feitelijk resultaat** | **Opmerking** |
| 1 | Open applicatie | Applicatie wordt geopend in de browser | Werkt | Gebruikte browser: Google Chrome Versie 74.0.3729.169 (Officiële build) (64-bits) |
| 2 | Klik op ‘Taxonomie’ | De pagina geeft het kopje ‘Taxonomie’ weer | Werkt |  |
| 3 | Vul een zoekterm in, in de tekstbalk ‘search’ | De zoekterm verschijnt in de zoekbalk | Werkt |  |
| 4 | Klik op ‘search’ met een ingevulde zoekterm | Resultaten die deze zoekterm bevatten worden getoond. | Werkt |  |
| 5 | Selecteer een attribuut via het dropdown menu ‘order by’ | Het gekozen attribuut wordt getoond. | Werkt |  |
| 6 | Klik op ‘asc<’ of ‘desc>’ | De gekozen radio-button wordt ‘gecheckt’ | Werkt |  |
| 7 | Klik op ‘search’ (met ingevulde attribuut en radiobutton) | De getoonde resultaten zijn volgens de gekozen orde en in juiste volgorde. | Werkt |  |

Tabel 2 Werking van 'Taxonomie'

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Stap** | **Instructie** | **Verwacht** | **Resultaat** | **Opmerking** |
| 1 | Open applicatie | Applicatie wordt geopend in de browser | Werkt | Gebruikte browser: Google Chrome Versie 74.0.3729.169 (Officiële build) (64-bits) |
| 2 | Klik op ‘Taxonomie’ | De pagina geeft het kopje ‘Taxonomie’ weer | Werkt |  |
| 3 | Type een zoekterm in | De zoekterm verschijnt in de zoekbalk | Werkt |  |
| 4 | Klik op search | Resultaten worden opgehaald uit de database en worden getoond. | Werkt grotendeels | Voor namen die vaker in de tabel voorkomen, maar wel een ander taxonomy ID hebben, is het niet mogelijk om de resultaten te tonen doordat de applicatie niet weet welk organisme bedoelt wordt |

Tabel 3 Werking van 'Zelf blasten'

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Stap** | **Instructie** | **Verwacht** | **Resultaat** | **Opmerking** |
| 1 | Open applicatie | Applicatie wordt geopend in de browser | Werkt | Gebruikte browser: Google Chrome Versie 74.0.3729.169 (Officiële build) (64-bits) |
| 2 | Klik op ‘Zelf blasten’ | De pagina geeft het kopje ‘Zelf blasten’ weer | Werkt |  |
| 3 | Type een sequentie in | De sequentie verschijnt in de zoekbalk | Werkt |  |
| 4 | Klik op blast | De sequentie wordt geblast en de resultaten worden getoond | Werkt niet | Het blasten geeft een error als het script wordt gerund in Azure, maar dit is niet het geval als het lokaal wordt gerund |

# Literatuurlijst

Altschul S, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J. *Mol. Biol*. 215:403-410. PubMed

Chapman B and Chang J: Biopython: Python tools for computational biology. *ACM SIGBIO Newsletter* 20 (2): 15–19 (August 2000)

Cock PA. Fields CJ, Goto N, Heuer ML, Rice PM (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*. 38 (6): 1767–1771

Copeland M, Soh J, Puca A, Manning M, Gollob D (2015). *Microsoft Azure: Planning, Deploying, and Managing Your Data Center in the Cloud*. Berkely, CA, USA: Apress.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016 Jan 4;44(D1):D733-45 PubMed

Oracle. (2019, 4 juni). *MySQL Connector/Python Developer Guide*. Geraadpleegd op 5 juni 2019, van <https://dev.mysql.com/doc/connector-python/en/>

Ravi RK, Walton K, Khosroheidari M. MiSeq: A Next Generation Sequencing Platform for Genomic Analysis. *Methods Mol Biol*. 2018;1706:223-232

Widenius M, Axmark D, DuBois P (2002). *Mysql Reference Manual.* Sebastopol, CA, USA: O’Reilly & Associates, Inc.

# Bijlagen

## Bijlage I Overige scripts

Scripts voor het automatisch blasten, het vullen van de database en het opvragen van de taxonomie zijn vooraf gebruikt om de data te genereren welke de applicatie aanroept. Deze scripts zijn terug te vinden via <https://github.com/BartJolink/Azure-Bi1a2>. De scripts zijn voorzien van commentaar om de werking hiervan te verduidelijken. In tabel 4 worden kort de scripts benoemd met een korte beschrijving.

Tabel 4 Benoeming en beschrijving overige scripts

|  |  |
| --- | --- |
| **Script (hyperlink)** | **Werking** |
| [Automatic BLASTx.py](https://github.com/BartJolink/Azure-Bi1a2/blob/master/Automatic%20BLASTx.py) | Automatisch ophalen van reads uit Excel bestand.  Reads blasten (BLASTx)  Resultaten parsen naar lijsten  Database vullen |
| [Automatic tBLASTx with check.py](https://github.com/BartJolink/Azure-Bi1a2/blob/master/Automatic%20tBLASTx%20with%20check.py) | Automatisch ophalen van reads uit Excel bestand.  Kijken welke reads nog niet in database zitten  Reads blasten (tBLASTx)  Resultaten parsen naar lijsten  Database vullen |
| [Taxanomy filler.py](https://github.com/BartJolink/Azure-Bi1a2/blob/master/Taxanomy%20filler.py) | Organisme naam ophalen uit database  Taxonomische rangschikking ophalen  Taxonomy ID’s (inclusief rang) in database zetten. |
| [Taxonomy ID Adder.py](https://github.com/BartJolink/Azure-Bi1a2/blob/master/Taxonomy%20ID%20Adder.py) | Organisme naam ophalen uit database  Taxonomy ID ophalen.  Taxonomy ID toevoegen aan blast\_resultaten tabel |