Materiaal en methode

Voor het onderzoek is er gebruik gemaakt van de sequentie data aangeleverd door het HAN BioCenter. Deze data is verkregen doormiddel van een metagenomics onderzoek, waarmee gebruik gemaakt is Illumina MiSeq technologie als paired-end reads en is gesequenst van DNA uit de compost die gebruikt wordt om champignons te kweken.

Voor het analyseren van de sequentie data is er de tool Blast gebruikt met als algoritme Blastx (Ian Korf, 2003-07-29) van NCBI (Benson, 2013). Er is gekozen voor Blastx om dat dit naar verwachting de meest nauwkeurige resultaten geeft binnen de tijd dat het onderzoek duurt.

Om de nauwkeurigheid van de kwaliteit van de alignments te verifiëren is er een scorematrix gebruikt. Hierbij werden de aanname gedaan dat de sequenties verwant zijn. BLOSUM62 is gebruikt omdat dit de beste balans is tussen de veruit elkaar gelegen sequenties en de dichterbij elkaar gelegen sequenties. Als database is de non-redundant GenBank gebruikt, eukaryoten zijn uitgesloten van de resultaten. Dit omdat dit onderzoek zich alleen gericht heeft op de prokaryoten en de mogelijk interessante eiwitten die zij produceren en daarmee van invloed kunnen zijn op de champignongroei. Voor elke gap in de sequentie is er 1 minpunt gerekend voor het openen van gap en 10 minpunten voor elke positie in de gap. Bacteriën muteren veel door hun hoge reproductiesnelheid waardoor er veel varianten ontstaan. Om deze reden is er iets soepeler geteld in het openen van een gap. De ‘word size’ is ingesteld op 6. ‘low complexity filter’ staat aan omdat repeats niet mee genomen worden in de Blast resultaten. Bacteriën bevatten geen tot nauwelijks repeats in hun genoom dus hier wordt geen rekening mee gehouden. Het opslaan van de gevonden matches doen we met een ‘cut-off’ of een E-value van 1\*10-10.

Alle gevonden matches met een lagere E-value zijn als te onnauwkeurig beschouwd en daarom niet meegenomen in het verdere onderzoek. Andere ‘cut-off’ criteria zijn een minimale identity van 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 de criteria zijn van belang voor een goede match.

Alleen de eerste 10 gevonden matches zijn opgeslagen in een Oracle database met behulp van data Modeler (Kolbe, April 2015) voor het maken van een ERD en SQL developer (Narayanan, Januari 2016) voor de opbouw van de database. De opzet van de database is als conceptueel ERD weergegeven in de bijlagen (Bijlagen 1). De database bestaat uit de kolommen: ‘Blast’ met hierin informatie over de uitgevoerde blast: date; description, ‘Omschrijving\_match’ met hierin informatie over de match: accessiecode; title; organisme; score; E\_value; identity; positives; gaps; frame; match en subject, ‘Sequentie’ met hierin informatie over de query sequentie: header; sequentie; ascii\_score. Verder is er een hiërarchie in de database verwerkt met familie, geslacht en soort.

Doormiddel van een programma geschreven in de programmeertaal Python (Rossum, 1995) en de modules BioPython (Peter J. A. Cock, march 2009), 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 match resultaten die voldoen aan de gevraagde query. Om de applicatie te testen zijn 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. Gebruikte imports zijn: van Bio.Blast NCBIWWW (Peter J. A. Cock, Module NCBIWWW, 2018) en NCBIXML (Peter J. A. Cock, Module Bio.Blast.NCBIXML, 2018), van de module Bio (Peter J. A. Cock, Package Bio, 2018) alleen Entrez (Peter J. A. Cock, Package Entrez, 2018) en er is gebruik gemaakt van een mysql connector (Oracle Corporation and/or its affiliates , 2018) voor de verbinding met de SQLdatabase.

Bijlage 1

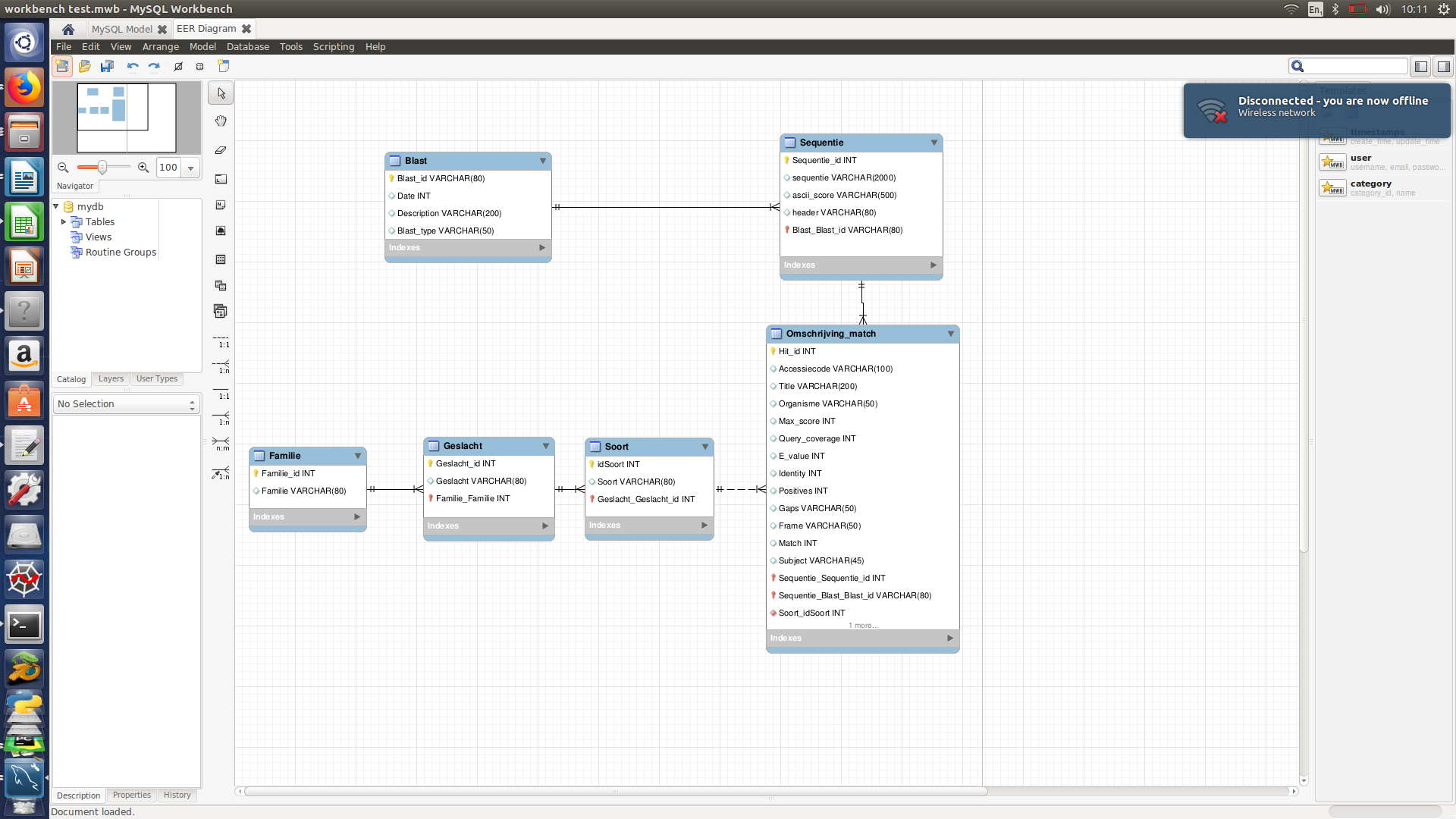


Figure Eigen werk