HAN, Nijmegen, Laan van scheut 2

Geautomatiseerde identificatie van micro-organismen

Valerie Verhalle BIN-1C
14-6-2018

Inhoud

SAMENVATTING
De champignon (Agaricus bisporus) is een schimmel uit de familie Agaricaceae. Bij de productie van champignons wordt speciale compost gebruikt. Een groot probleem binnen de champignonteelt is echter dat de oogst van champignons sterk kan variëren. Er ontbreken op dit moment methoden om aanwezige micro-organismen in de compost te detecteren, karakteriseren en te kwantificeren. Bij het gebruik van NGS data is een juiste annotatie van de reads van groot belang. Voor het ontwikkelen en testen van een betrouwbare annotatie methode is een specifieke dataset gegenereerd. Op verschillende momenten tijdens een commercieel composteringsproces is een compost sample genomen en daaruit is het DNA geïsoleerd. Het DNA is vervolgens gesequenced door het bedrijf BaseClear met behulp van de Illumina MiSeq technologie als paired-end reads (per sample ongeveer 2500000 reads). Het doel van dit onderzoek is om 200 sequenties uit de verkregen dataset van annotatie te voorzien d.m.v. een webapplicatie die sequenties geautomatiseerd kan blasten met behulp van de modules Blastx en BioPython. De verkregen annotaties worden door de applicatie in een SQL database geladen. Wat opviel was dat er van 107 organismen geen soort bekend is. Verder viel op dat, op een paar uitzonderingen na, er weinig organismen dubbel voorkwamen. Door de 200 sequenties van een annotatie te voorzien heeft dit onderzoek er voor gezorgd dat er een beter inzicht is in de microflora van de compost afkomstig van champignonkwekers. Een beperking van het onderzoek is dat er maar 200 sequenties van een annotatie zijn voorzien, beter zou zijn geweest als alle 2500000 reads van een annotatie waren voorzien. Een mooi vervolg onderzoek zou kunnen zijn: ‘Welke van de geannoteerde sequenties hebben een negatieve of positieve invloed op de champignongroei?’ of ‘Kunnen we voorspellen welke microflora zich zal ontwikkelen in de verschillende stadia van de champignongroei?’.

INLEIDING

Bij de productie van champignons wordt een mengsel van onder andere paarden-, kippenmest, stro en gips gecomposteerd en vervolgens beënt met mycelium van de champignon *Agaricus bisporus*. De beënte compost wordt in vrachtwagens naar de champignonkwekers overgebracht waar onder goed gecontroleerde omstandigheden de champignons worden geteeld en geoogst. Het bereidingsproces van de compost is er op gericht om een optimale voedingsbodem voor de champignon te verkrijgen en daarnaast de groei van ongewenste micro-organismen te voorkomen (zie [Straatsma et al., 1994]). Een groot probleem binnen de champignonteelt is echter dat de oogst van champignons sterk kan variëren. De in de compost aanwezige micro-organismen die te samen de microflora vormen spelen daarbij een grote rol. Tijdens de compost bereiding wisselt de samenstelling van deze microflora voortdurend en heeft invloed op de afbraak van de compost (fermentatie) en de opbouw van voor de champignon geschikte voeding. Tevens zal de microflora een positieve of negatieve rol spelen richting micro-organismen die schadelijk zijn voor de champignonteelt (pathogenen en onkruidschimmels) [Largeteau and Savoie, 2010].

De afgelopen jaren zijn er verschillende studies gepubliceerd met betrekking tot de detectie, karakterisering en kwantificatie van de aanwezige micro -organismen in de compost (bijv: [Silva et al., 2009]). Echter, de meeste van deze studies beschreven de aanwezige microbiële populatie in meer algemene zin. Met de opkomst van de nieuwe Next Generation Sequencing methoden is het mogelijk geworden de populatie veel nauwkeuriger te beschrijven tegen lage kosten [Oulas et al. 2015]. Met dit onderzoek hopen we antwoord te verkrijgen op de vragen: ‘welke micro-organismen kun je identificeren aan de hand van de dataset?’ en ‘welke eiwitten kun je identificeren aan de hand van de dataset?’. De analyse van het zogenaamde 'Metagenoom' vindt routinematig op twee manieren plaats. De meest gebruikte methode is metagenomics op grond van specifieke marker genen, meestal 16S ribosomaal RNA [Tringe and Hugenholtz, 2008]. Hierbij wordt DNA geïsoleerd uit een bepaalde omgeving van interesse en het 16S rRNA coderende deel met a-specifieke primers geamplificeerd [Klindworth et al., 2013] en vervolgens gesequenced. Met behulp van de 16S rRNA sequenties aanwezig in referentiedatabases zoals bijv GreenGenes [DeSantis et al., 2006] kunnen de resulterende ‘reads’ dan worden geannoteerd, d.w.z. voorzien van een organisme naam en een plaats in de taxonomie. De tweede metagenomics methode spitst zich toe op de karakterisering van alle aanwezige genen. Hier wordt alle aanwezige DNA gesequenced en worden de ‘reads’ functioneel geannoteerd met tools zoals bijv MG-RAST [Meyer et al., 2008]. Beide typen analyse leveren aanvullende informatie. De 16S rRNA analyse onder verschillende condities geeft een beeld over de verschuiving die plaats vindt in de soorten samenstelling binnen een populatie, terwijl de volledige DNA analyse een beeld geeft van de verschuiving die plaats vindt in de genen samenstelling van de populatie. Zelfs bij gelijkblijvende soort-samenstelling kan de gen-samenstelling veranderen wanneer er een verschuiving plaats vindt in de aanwezige stammen.

Om de effecten van de samenstelling van de microflora in compost op de uiteindelijke opbrengst aan champignons te kunnen bestuderen moet de kwalitatieve en kwantitatieve samenstelling van de microflora in de verschillende fases van het composteringsproces betrouwbaar kunnen worden bepaald. Bij het gebruik van NGS data is daarvoor een juiste annotatie van de reads van groot belang. Voor het ontwikkelen en testen van een betrouwbare annotatie methode is een specifieke dataset gegenereerd. Op verschillende momenten tijdens een commercieel composteringsproces is een compost sample genomen en daaruit is het DNA geïsoleerd. Het DNA is vervolgens gesequenced door het bedrijf BaseClear met behulp van de Illumina MiSeq technologie als paired-end reads (per sample ongeveer 2500000 reads). Voor het opzetten en testen van een annotatie-pipeline is een kleine test set reads genomen (~100 sequenties) en deze zijn middels een BLAST search van een functie annotatie voorzien. Hierbij werden de volgende aannames gemaakt: in de compost van de champignonkweker komen alleen micro-organismen voor die sequenties in hun genoom hebben. Deze sequenties coderen voor eiwitten die van invloed kunnen zijn op de champignonteelt. Op deze manier konden 200 sequenties van een specifieke annotatie worden voorzien. Daarbij viel op dat de soort Paenibacillus 39 keer voorkwam en het organisme Paenibacillus amylolyticus maar 5 keer. Blijkbaar komt de soort Paenibacillus vaker met verschillende families voor. De verkregen annotaties werden vergeleken met de non-redundant GenBank en daarbij viel op dat van 107 organismen de soort onbekend is en van 31 organismen het geslacht onbekend is. De gebruikte methode is wel eenvoudig op te schalen, zodat ook de complete datasets van 2500000 reads relatief snel en betrouwbaar kunnen worden geannoteerd.

MATERIAAL EN METHODEN

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 gesequenced 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). Voor het blasten zijn de volgende parameters gebruikt: module: Blastx, scorematrix: BLOSUM62 en de database: non-redundant GenBank. 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. 'word size' is ingesteld op 6, 'low complexity' filter staat aan, E-value van $1*10^{-10}$, identity van 25% en hoger, positive percent van miniaal 70% en een query coverage hoger dan 80%.

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, Biopython: freely available Python tools for computational molecular biology and bioinformatics, 1st edition, 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.

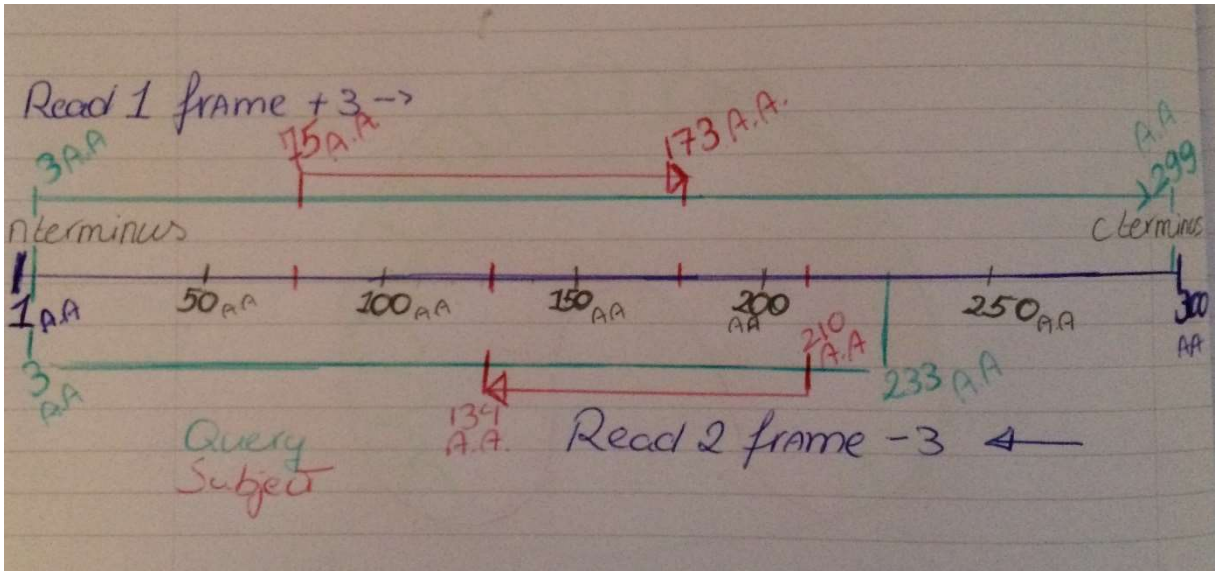
De website laat de gebruiker de verschillende micro-organismen opvragen, de Blast resultaten en match resultaten die voldoen aan de gevraagde query. Om de applicatie te testen zijn er 10 forward en 10 reverse 'reads' geautomatiseerd geblast en vervolgens via een ander script automatisch ingeladen in 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.

RESULTATEN

Het doel was om 200 sequenties van annotatie te voorzien door middel van een applicatie die de 200 sequenties geautomatiseerd kan blasten en ze vervolgens in een database kan laden. Deze applicatie kan geraadpleegd worden via een site waarmee de database doorzocht kan worden. Eventueel kunnen blast resultaten die via de site zijn verkregen ook worden ingeladen in de database. Het uiteindelijke doel van dit onderzoek was om de champignonteelt te stabiliseren door meer inzicht te krijgen in welke micro-organismen zich in de verschillende stadia van de compost bevinden. Een van deze micro-organismen met geproduceerd eiwit wordt verder toegelicht, zoals dit in het verdere stadium van dit onderzoek gedaan zou kunnen worden met alle geannoteerde sequenties.

Uit de dataset zijn 222 organismen en 222 eiwitten geïdentificeerd. De 3 meest voorkomende organsimen zijn: *Paenibacillus amylolyticus* (komt 5 keer voor), *Methanosarcina mazei* (komt 4 keer voor) en *Treponema bryantii* (komt 3 keer voor). De 3 meest voorkomende eiwitten zijn: molybdenum cofactor *biosynthesis protein B* (komt 42 keer voor), *sugar ABC transporter permease* (komt 42 keer voor) en *purine-nucleoside phosphorylase* (komt 42 keer voor).

Het organisme waar verder op ingegaan zal worden is een van de 200 sequenties dat door Illumina MiSeq technologie als paired-end reads is gesequenced, vervolgens geblast met de module Blastx en zo van annotatie voorzien. In de database is te zien dat het organisme *Actinotalea fermentans ATCC 43279 = JCM 9966 = DSM 3133 (accession: KGM15509)* regelmatig voorkomt. In de annotatie is te zien dat de blast resultaten een hit hebben gegeven op het eiwit dat dit organisme produceert: 2-isopropylmalate synthase, geproduceerd door het leuA gen. In het onderstaande figuur (figuur 1) is te zien hoe de forward en reverse read op de sequentie van de gevonden hit liggen. Met in het groen de query en in het rood het subject.



Figuur 1 Overzicht van de forward en reverse read, hoe deze ten opzichte van de sequentie van de hit liggen.

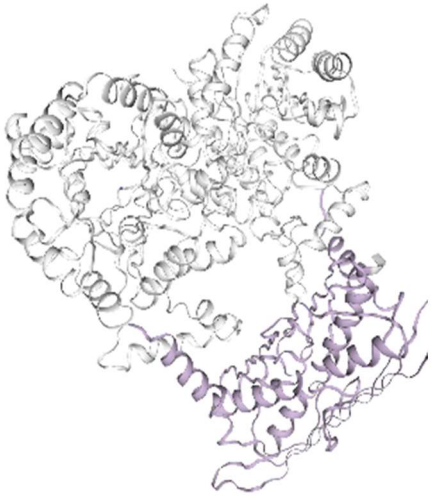
Het organisme *Actinotalea fermentans* behoort tot het rijk van de bacteriën en is onderverdeeld in de stam van *terrabacteriën*, die er om bekend staan dat ze goed bestand zijn tegen milieugevaren zoals uitdroging, ultraviolette straling en hoge zoutgehaltes. Verder behoort de bacterie tot de klasse van de *actinobacteria*, de order van de *micrococcineae* en de familie van de *cellulomonadaceae*, die behoren tot de groep grampositieve bodemorganismen (Nightingale A, 2017). Geslacht en soort zijn van de naam af te leiden. De cellular component van deze bacterie beslaat alle structuren van een cel uitgesloten van het plasma membraan en de nucleus. De bacterie is zo goed als onschadelijk voor een gezonde volwassene en kan in een BioHazard level 1 laboratorium gehouden worden. *Actinotalea fermentans* is een aeroob organisme wat inhoud dat het zuurstof nodig heeft voor zijn energie productie. Zuurstof wordt gebruikt in cellulaire processen zoals de Kreb's cycle en de elektronen transport om ATP te produceren. Wanneer de zuurstofatomen niet meer nodig zijn worden er steeds twee zuurstofatomen aan een carbonatoom gebonden, zodat het makkelijk als CO_2 vervoerd kan worden. Het Organisme heeft een maximale activiteit bij 45 graden Celcius en bij een Ph van 8.

Het eiwit *2-isopropylmalate synthase* heeft 2 GO termen: het 'Biologisch proces' en de 'Moleculaire functie'. Het biologische proces van het eiwit resulteert in de chemische reactie die leidt tot de vorming van het aminozuur leucine: 2-amino-4-methylpentanoic acid(leuA). *2-isopropylmalate synthase* is niet de enige katalysator, er zijn meerdere proteïnes betrokken bij het omzetten van 3-methyl-2-oxobutanoate + acetyl-CoA naar leucine. De volgorde is parallel met de L-leucine biosynthesis pathway: *3-isopropylmalate dehydratase small subunit (leuD), 3-isopropylmalate dehydratase large subunit (leuC), 3-isopropylmalate dehydrogenase (leuB) en Branched-chain-amino-acid aminotransferase (Accession: N867_09825). 2-isopropylmalate synthase* wordt door de ribosomen van de cel geproduceerd.

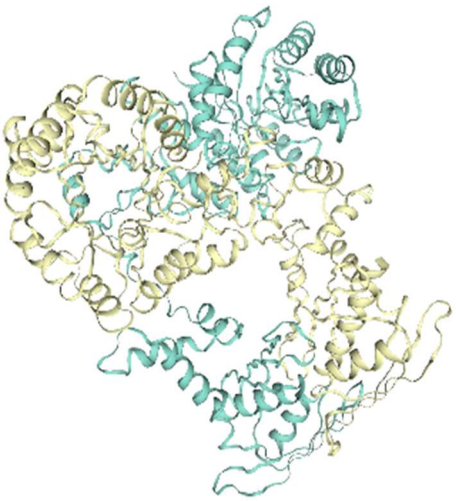
Het eiwit *2-isopropylmalate synthase* is de 1e katalysator in de L-leucine biosynthesis pathway. De moleculaire functie van het eiwit is het katalyseren van de eerste stap in de L-leucine biosynthesis pathway zoals in de reactievergelijking (een aldol condensatie) is weergegeven:

3-methyl-2-oxobutanoate + acetyl-CoA + H_2O →(2S)-2-isopropylmalate + CoA + H^+

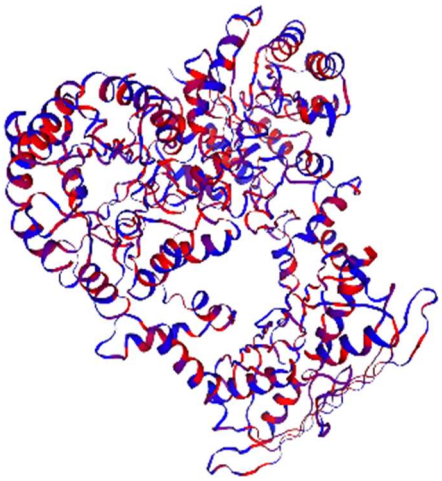
In de bijlagen is een figuur weergegeven met daarin de complete reactie van *2-isopropylmalate synthase* naar leucine (Bijlage2) en de GO termen die horen bij deze reactie(Bijlage3). In figuur 3 is te zien dat *2-isopropylmalate synthase* uit 2 unieke groepen bestaat (geel en groen), wat opvalt is dat de gele chain een homo2-mer is en de groene chain ook uit 2 keer dezelfde polypeptiden bestaat. Dit is logisch want *2-isopropylmalate synthase* heeft een subunit structure genaamd homotetramer (Chen F., 2017). In figuur 4 is de hydrofobiciteit van het eiwit te zien.



Figuur 2. 2-isopropylmalate synthase LeuA, alloste ric (dimerisation) domain



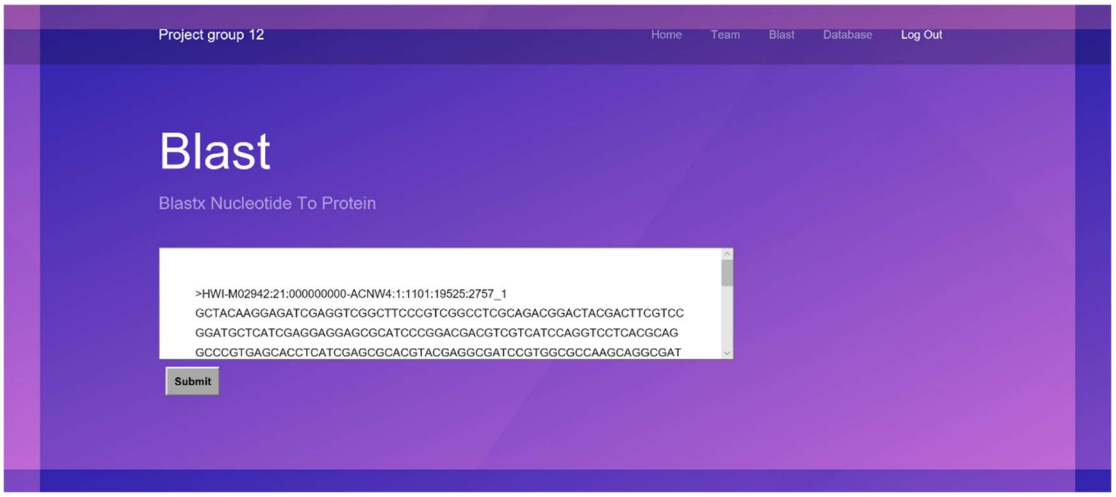
Figuur 3. 2-isopropylmalate synthase



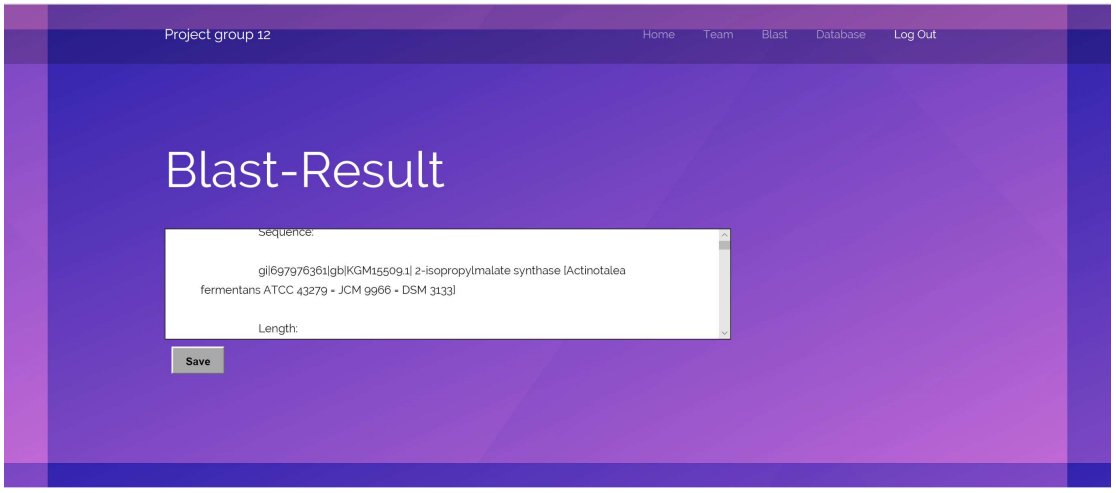
Figuur 4. 2-isopropylmalate synthase hydrofobic. Met in het blauw de hydrofiele domeinen en in het rood de hydrofobe domeinen.

Het gebruikte python script is onderverdeeld in 6 functies. De eerste functie is zo geschreven dat het een txt file met ruwe data ophaalt en dit omzet naar een overzichtelijk fasta file. Dit fasta file wordt vervolgens weggeschreven en kan door de volgende functie weer aangeroepen worden. Parameters voor het blasten worden aan de module NCBIWWW.qblast meegeven en de resultaten van deze blast worden vervolgens opgeslagen in een variabele, die wordt weggeschreven als xml file. Wat door volgende functies makkelijk kan worden aangeroepen. De benodigde informatie (hit_id, organisme, accessioncode, title, score, E-value, identities, positives, gaps, frame, match en subject) wordt uit het xml file opgevraagd en via een SQL query ingeladen naar de SQL database. Voor de taxonomie wordt per hit de accessiecode opgevraagd uit het eerder gegenereerde xml file. Door de accessiecode mee te geven aan de module Entrez kan de taxonomie van de hit opgehaald worden. De taxonomie wordt opgesplitst in de variabelen: Familie, Geslacht en Soort, deze worden dan ook via een SQL query ingeladen in de SQL database. Dit python script is geschreven om de 200 sequenties makkelijk van een annotatie te kunnen voorzien en ze gemakkelijk in de SQL database te laden. Een zelfde principe is gebruikt in de webapplicatie FLASK maar hier is het de bedoeling dat de gebruiker van de site een(of meerdere) sequenties blast en ze daarna eventueel update naar de database.

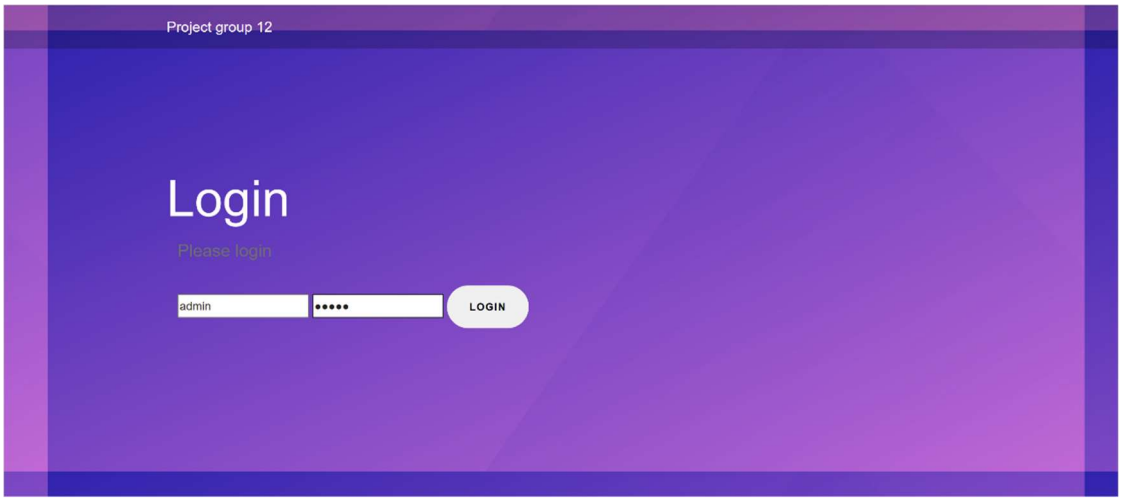
In de webapplicatie FLASK wordt er gebruik gemaakt van een zelfde soort script, alleen is er dan al een fasta-input aanwezig (figuur 5). De fasta-input moet voorzien zijn van een header en de sequentie moet een nucleotide sequentie zijn. Met vastgestelde parameters zal de input geblast worden tegen de NR-database. De output wordt op een nieuwe pagina overzichtelijk weergegeven (figuur 6). Nadat de Blastoutput is weergegeven kan de gebruiker er voor kiezen om de gegevens in de SQL database te laden(Bijlage 1). Om er voor te zorgen dat hier geen misbruik van gemaakt wordt is er voor gekozen om een inlogscherm(Figuur 7) te gebruiken. De gebruiker zal de loginnaam en het wachtwoord eerst correct moeten invullen voordat er gebruik kan worden gemaakt van de website.



Figuur 5 Blast pagina, input is een nucleotide sequentie die met Blastx geblast wordt tegen de NR-database.



Figuur 6 Blast-Result, op deze pagina worden de blast resultaten overzichtelijk weergegeven. Mochten de resultaten interessant zijn voor het onderzoek, dan kunnen de resultaten worden ingeladen in de SQL database door op 'save' te klikken.



Figuur 7 Inlogscherm, wanneer iemand gebruik wil maken de site zal hij/zij eerst een correct wachtwoord moeten invullen. Dit om misbruik te voorkomen.

DISCUSSIE

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 gesequenced van DNA uit de compost die gebruikt wordt om champignons te kweken.

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.

Lang niet alle sequenties waren van goede kwaliteit, de meeste hadden erg slechte ascii-scores. Voornamelijk de reverse reads waren slecht gesequenced. Desondanks zijn de 200 sequenties van een annotatie voorzien, er zullen dus annotaties tussen zitten die niet accuraat zijn.

Meerdere blast resultaten hebben een hit gegeven op het 2-isopropylmalate synthase eiwit geproduceerd door het organisme *Actinotalea fermentans* ATCC 43279 = JCM 9966 = DSM 3133 (accession: KGM15509). Het eiwit wordt door de ribosomen van de cel geproduceerd en beschikt over twee GO termen: het 'Biologisch proces' en de 'moleculaire functie'. Het biologisch proces van dit eiwit is het in gang brengen van de chemische reactie die leidt tot het aminozuur leucine. De moleculaire functie is het katalyseren van de eerste stap in de L-leucine biosynthesis pathway (bijlage 1). Van het eiwit 2-isopropylmalate synthase was geen 3d structuur aanwezig. Daarom is er voor gekozen om een ander organisme toe zoeken wat ook het 2-isopropylmalate synthase eiwit produceert, waar wellicht wel een 3d structuur van aanwezig is. Het organisme *Leifsonia xyli subsp. xyli (strain CTCB07)* bezit ook het leuA gen en produceert daarmee ook het 2-isopropylmalate synthase eiwit, hier waren wel 3d structuren aanwezig.

Uit de dataset zijn 222 eiwitten geïdentificeerd, van deze eiwitten kwamen de volgende 3 het vaakst voor: molybdenum cofactor *biosynthesis protein B* (komt 42 keer voor), *sugar ABC transporter permease* (komt 42 keer voor) en *purine-nucleoside phosphorylase* (komt 42 keer voor). Het *sugar ABC transporter permease* eiwit is gekarakteriseerd door twee nucleotide-binding domein (NBD) en twee transmembraan domeinen (TMDs) (Wilkens, 2015). ABC transporters zijn van belang voor alle vormen van leven om, deze reden is het niet raar dit eiwit zich in het sample van de compost bevindt.

Verder zijn er ook 222 organismen geïdentificeerd. *Paenibacillus amylolyticus* kwam van de 222 organismen 5 keer voor. Veel *Paenibacillus* soorten ondersteunen de groei van gewassen door nitrogeen fixatie, fosfaat solubilisatie en het beschermen van de gewassen tegen herbivoren insecten en pathogenen bacteriën, fungi en virussen (Elliot Nicholas Grady, 2016). Ook van dit organisme is het niet opvallend dat het in de dataset voorkomt.

Door de 200 sequenties van een annotatie te voorzien heeft dit onderzoek er voor gezorgd dat er een beter inzicht is in de microflora van de compost afkomstig van champignonkwekers. Dit houdt in dat onderzoekers nu meer zeker zijn van de microflora in de compost. Een beperking van het onderzoek is dat er maar 200 sequenties van een annotatie zijn voorzien, beter zou zijn geweest als alle 2500000 reads van een annotatie waren voorzien. Een mooi vervolg onderzoek zou kunnen zijn: 'Welke van de geannoteerde sequenties hebben een negatieve of positieve invloed op de champignongroei?' of 'Kunnen we voorspellen welke microflora zich zal ontwikkelen in de verschillende stadia van de champignongroei?'.

De onderzoeksvragen die aan het begin van dit onderzoek gesteld werden waren: ‘Welke micro-organismen kun je identificeren aan de hand van de dataset?’ en ‘Welke eiwitten kun je identificeren aan de hand van de dataset?’. Aan de hand van de dataset hebben we onder andere de bacterie *Actinotalea fermentans* ATCC 43279 = JCM 9966 = DSM 3133 (accession: KGM15509) kunnen identificeren. Deze bacterie produceert onder andere het eiwit het 2-isopropylmalate synthase, dat verantwoordelijk is voor het katalyseren van de chemische reactie van 3-methyl-2-oxobutanoate + acetyl-CoA naar L-leucine.

Bibliografie

Benson, D. A.-M. (2013). *Nucleic Acids Research*. Opgehaald van NCBI; PMC: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531190/

Chen F., L. Y. (2017, october). *UniProtKB - A0A0A0C873 (A0A0A0C873_9CELL)*. Opgehaald van UniProt: http://www.uniprot.org/uniprot/A0A0A0C873#similar_proteins

Elliot Nicholas Grady, J. M.-C. (2016, december 1). *Current knowledge and perspectives of Paenibacillus: a revies*. Opgehaald van PMC: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5134293/

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.

Nightingale A, A. R. (2017). *Taxonomy - Actinotalea fermentans ATCC 43279 = JCM 9966 = DSM 3133* . Opgehaald van UniProt: http://www.uniprot.org/taxonomy/862422

Oracle Corporation and/or its affiliates . (2018). *MySQL Connector/Python Developer Guide*. Opgehaald van MySQL: https://dev.mysql.com/doc/connector-python/en/

Peter J. A. Cock, T. A. (2018, april 4). *Module Bio.Blast.NCBIXML*. Opgehaald van Package Bio: Package Blast: Module NCBIXML: http://biopython.org/DIST/docs/api/Bio.Blast.NCBIXML-pysrc.html

Peter J. A. Cock, T. A. (2018, april 4). *Module NCBIWWW*. Opgehaald van Package Bio; Package Blast: Module NCBIWWW: http://biopython.org/DIST/docs/api/Bio.Blast.NCBIWWW-module.html

Peter J. A. Cock, T. A. (2018, april 4). *Package Bio*. Opgehaald van Package Bio: https://biopython.org/DIST/docs/api/Bio-module.html

Peter J. A. Cock, T. A. (2018, april 4). *Package Entrez*. Opgehaald van Pachage Bio: Package Entrez: https://biopython.org/DIST/docs/api/Bio.Entrez-module.html

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).

Wilkens, S. (2015, februari 3). *Structure and mechanism of ABS transporters*. Opgehaald van PMC: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4338842/

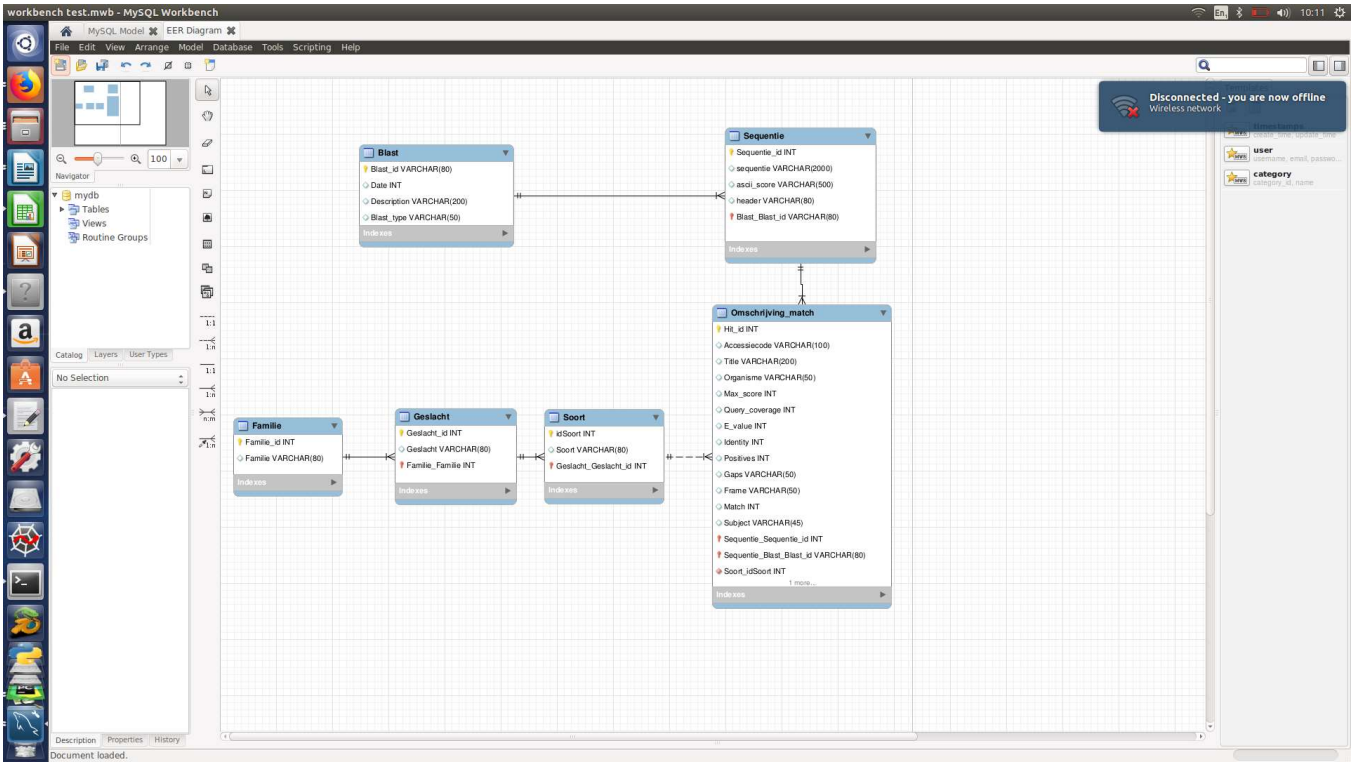
Willems, K. (2017, February 22). *Matplotlib Turtorial: Python Plotting*. Opgehaald van DataCamp: https://www.datacamp.com/community/tutorials/matplotlib-tutorial-python

- DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol. 2006 72:5069-72. http://greengenes.lbl.gov/cgi-bin/nph-index.cgi
- Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies.Nucleic Acids Res. 2013 41(1):e1.
- Largeteau ML, Savoie JM. Microbially induced diseases of *Agaricus bisporus*: biochemical mechanisms and impact on commercial mushroom production. Appl Microbiol Biotechnol. 2010 86(1):63-73.
- Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 2008 9:386.
- Oulas A, Pavloudi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N, Kotoulas G, Arvanitidis C, Iliopoulos I. Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. Bioinform Biol Insights. 2015 9:75-88
- Silva CF, Azevedo RS, Braga C, da Silva R, Dias ES, Schwan RF. Microbial diversity in a bagasse-based compost prepared for the production of *Agaricus brasiliensis*. Braz J Microbiol. 2009 40(3):590-600.
- Straatsma G, Olijnsma TW, Gerrits JP, Amsing JG, Op Den Camp HJ, Van Griensven LJ. Inoculation of *Scytalidium thermophilum* in Button Mushroom Compost and Its Effect on Yield. Appl Environ Microbiol. 1994 60(9):3049-3054.

  Tringe SG, Hugenholtz P. A renaissance for the pioneering 16S rRNA gene. Curr Opin Microbiol. 2008 11(5):442-446.

BIJLAGEN

Bijlage 1



Figuur 8. Eigen werk, ERD van de gebruikte SQL databse

Bijlage 2

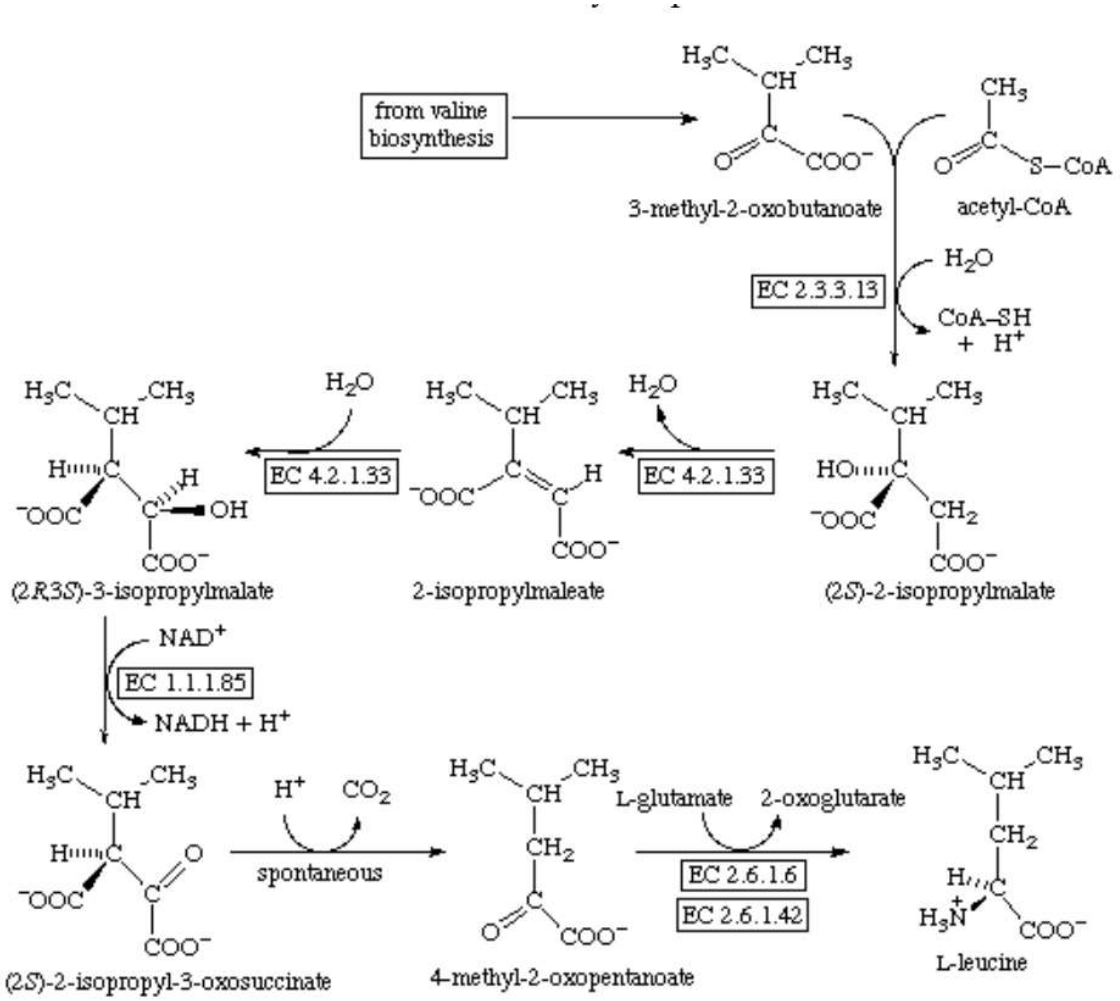


Figure 9 De chemische reactie tussen 2-isopropylmalate synthase naar L- leucine.
http://www.sbcs.qmul.ac.uk/iubmb/enzyme/reaction/AminoAcid/Leu.html

Bijlage 3

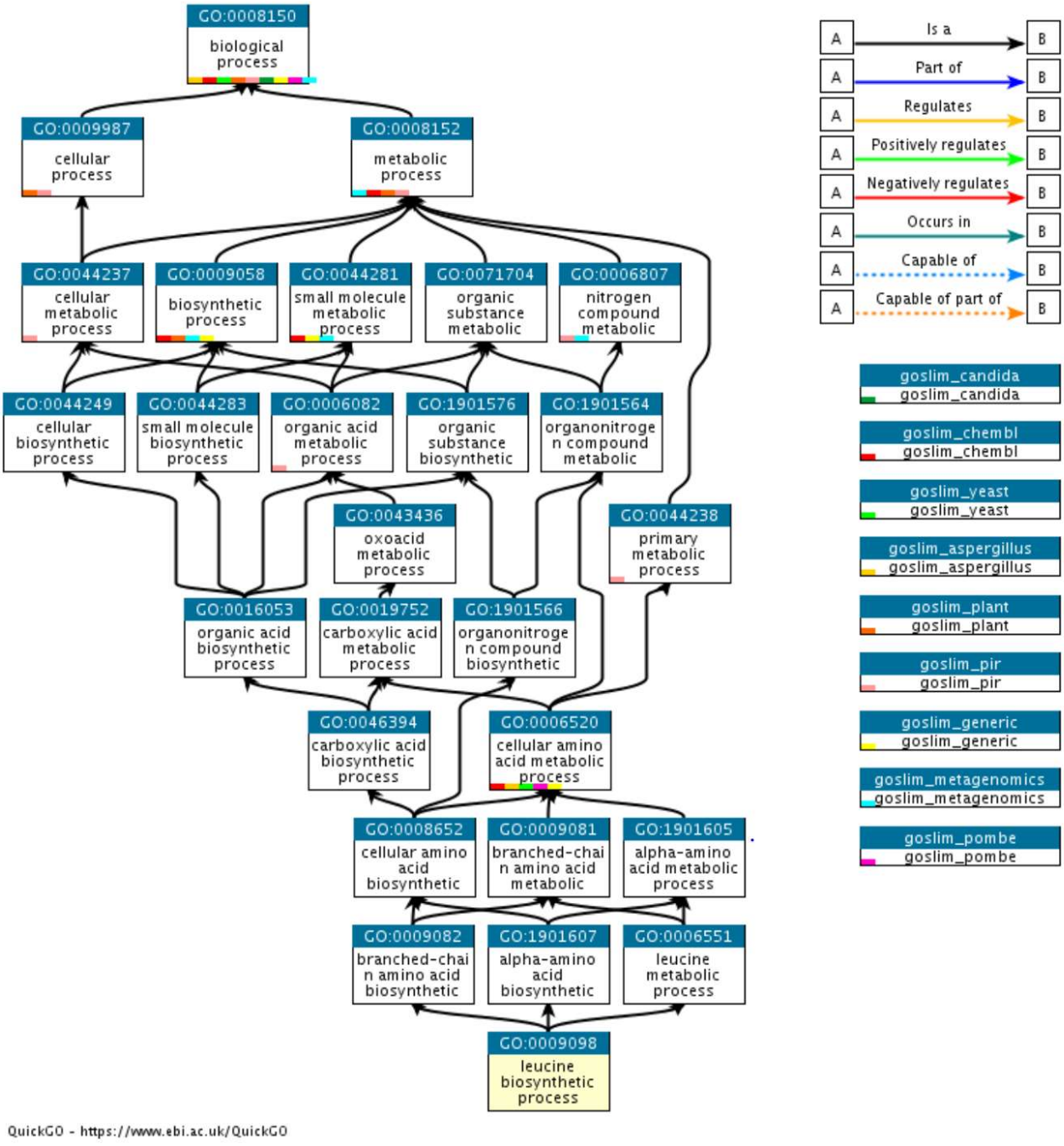


Figure 10 GO termen bij de vorming van L-Leucine d.m.v. een chemische reactie met 2-isopropylmalate synthase. https://www.ebi.ac.uk/QuickGO/term/GO:0009098