# Projectopdracht

## Projectomgeving

Genetic mapping is een methode voor het onderzoeken van unieke genetische patronen die gekoppeld kunnen worden aan een fysieke eigenschap. Moleculaire markers worden gemapped tegen het genoom van een populatie. Het resultaat is een verzameling van quantative trait loci (QTL), dit zijn fragmenten van het dna die samen voorkomen met een bepaald fenotype.

Op deze manier zijn ook de QTL’s en vervolgens de genen bepaald voor anthocyaan productie door modelorganisme Arabidopsis Thaliana. Anthocyanen behoren tot de flavonoiden en zijn stoffen die te vinden zijn in de vacuole van verschillende hogere planten. Ze produceren een karakteristieke rode of paarse kleur en hebben een aantal belangrijke functies waaronder: bescherming tegen droogte, zoute omgeving, te veel zonlicht en insecten.

De experimentele methode voor het vinden van genen die een rol spelen bij een specifiek biologisch proces is vaak niet voldoende voor het verkrijgen van een compleet beeld. Daarom wordt er gezocht naar een aanvullende methode voor het identificeren van betrokkenen genen. Daarnaast is er niet voldoende bekend over het effect van anthocyanen in relatie tot de omgeving en de plantensoort waarin ze voorkomen.

Doelstelling

Het doel is een completer beeld te krijgen van functies van anthocyanen, welke andere biologische processen (genen) erbij betrokken zijn en hoe dit zich verhoudt tot de plantensoort en de omgeving waarin ze voorkomen.

Opdrachtformulering

1. Welke planten produceren anthocyanen in stress omstandigheden?
2. Welke genen (orthologen) zijn betrokken bij de anthocyanen productie in verschillende plantensoorten?
3. Welke omstandigheden beinvloedt de productie van anthocyanen?

cooccurence (samen voorkomen) van verschillende genen bij een bepaald fenotype.

Het onderzoek naar cooccurence is in tegenstelling tot gene mapping niet experimenteel maar volledig in silico. Met behulp van textmining (geautomatiseerd scannen van teksten op bepaalde begrippen) van relevante wetenschappelijke artikelen kan er worden bepaald in welke mate bepaalde gennamen samen worden vernoemd. Hieruit wordt een aanname gemaakt dat de cooccurence in de tekst duidt op een bepaald biologisch verband.

## Op te leveren producten en diensten

Database

Een SQL database waarin de alle termen opgenomen met daarbij de link naar een andere term en optelsom van de cooccurence van deze termen. Deze informatie is het minimale van wat er kan worden opgeslagen. Afhankelijk van de textmining library die uiteindelijk zal worden geïmplementeerd zal deze database worden uitgebreid met de data waar het algoritme van het programma van afhankelijk is.

Textmining

Er zijn verschillend Open Scource libraries voor het toepassen van texmining waaronder, OpenNLP (apache natural language processing, erg uitgebreid, machine learning implementatie is mogelijk, goede documentatie), Lingpipe (vooropgezette database, voor niet-experts, goed documentatie) en JTMT (Java Text Mining Tool, kleiner eenmansproject, documentatie is geringer).

Visualisatie

Het eindproduct is een visualisatie van de cooccurence van verschillende begrippen of een verzameling van woorden die voorkomen in teksten betreft anthocyanen. De begrippen zijn relevant aan de onderzoeksvragen en zullen oa bestaan uit: gennamen, fenotypen, milieu omschrijvingen, soortnamen, biologische processen (GO termen). De visualisatie maakt dus verbanden duidelijk zonder dat de gebruiker aangeeft tussen welke twee (of meer) begrippen er een verband moet worden gezocht. De visualisatie is een product van een geautomatiseerd programma dat textmining toepast op wetenschappelijke artikelen en eventueel andere teksten die omschrijving bevatten van een onderzoek naar anthocyanen.

De visualisatie kan handmatig worden opgebouwd met JavaFX. Een ander mogelijkheid is een JavaScript API tool met vooropgezette modellen zoals FusionCharts, JavaScript InfoVis Toolkit, Zingchart of Chart.js.