

30 april 2017

Koen van der heide, Thomas Reinders en rick beeloo

[Bedrijfsnaam]

[Bedrijfsadres]

Plan Van aanpak

course 8a

***Introduction***

Anthocyanins are water-soluble vacuolar pigments that may appear in various colors depending on the pH. Besides coloration they play an important role in plants against reactive oxygen species caused by abiotic stresses such as overexposure to ultra-violet light and temperature. They also have potential in food value. Anthocyanins belong to a parent class of molecules called flavonoids synthesized via the phenylpropanoid pathway. The diversity of properties lead to a general question introduced by Anna Czerednik: “How to improve plants using secondary metabolites (anthocyanins)” This question was broken down into three more manageable questions: “Which plants produce anthocyanin(s)?”,”which genes regulate anthocyanin production?” and “Which circumstances effect anthocyanin production?”. Translating these questions in bioinformatics questions: Which genes are coupled to the term “anthocyanin” in the NCBI gene database? to which organisms do these genes correspond? What conditions are mentioned in published articles concerning the previously found genes that have an influence on the expression of these genes? Our approach to answer these questions will consist of database searches in combination with text-mining.

***Material & methods***

We will determine which plants produce anthocyanin and which conditions might influence this production by first searching for genes known to be involved in anthocyanin production in general. The names of the genes found, combined with certain keywords indicating a conditional change, will then be searched in biomedical research articles (PubMed database1) to discover known relationships between certain conditions and changes in gene expression levels.

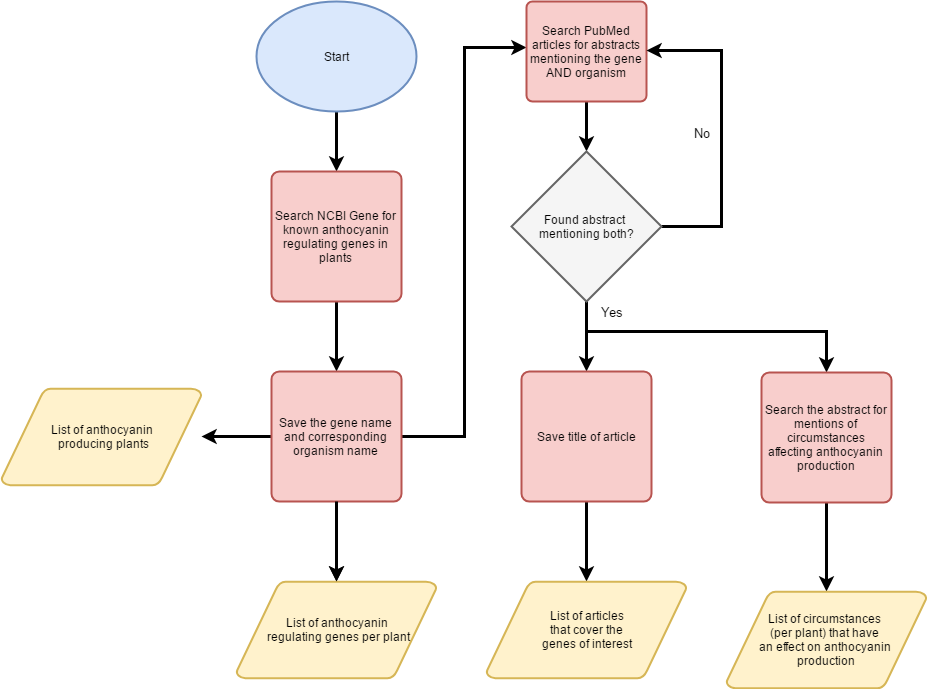


Figure1, **Flowchart**. This flowchart shows the different steps which will answer three questions: “Which plants produce anthocyanin(s)?”,”which genes regulate anthocyanin production?”, “Which circumstances effect anthocyanin production?”. A more detailed overview of “Search the abstract for mentions of circumstances affecting anthocyanin production” can be seen in **figure 2**.

Retrieval of anthocyanin genes

The NCBI Gene (Maglott, Ostell, Pruitt, & Tatusova, 2010) database will serve as starting point to find anthocyanin coding genes (or genes which influence the expression of anthocyanin coding genes). To retrieve these genes the NCBI database will be queried using “anthocyanin” accompanied by a plant filter, or more specific a filter for “Embryophyta”. This will result in a list of genes. Fortunately NCBI couples every gene to the organism in which it was found. Therefore the second research question: “Which plants produce anthocyanin?” can be directly inferred from the gene list (assuming that if a gene for anthocyanin is present, this plant will be either naturally or artificially capable to express this gene). Offcourse this is not the only approach. Another approach would be the usage of text-mining. However we prefer to not use text-mining in this case. The first and most important reason is reliability, despite the advanced training methods a text mining algorithm will never fit every article. Basically these training approaches build a model from the input data and use this to search words/sentences which fit this model. However there is no “standard” for writing articles, as a result some authors will use deviating words and/or deviating sentence structures, making it inevitable to miss information. Secondly search time should be considered. The NCBI gene database has predefined keywords coupled to every entry (e.g. anthocyanin in this case). Consequently a search against this database is extremely fast whereas text-mining searches are proportional to the inputted text (thus the number of articles). Lastly text-mining is limited in the sense of data availability. These limitations are either due to PubMed or the article authors. For example PubMed only allows the retrieval of article titles, abstracts and keywords (due to transfer size). Even if we bypass this and download full articles we will be hindered by articles which are not freely accessible and significantly increase the search time (as stated above). On the other hand the NCBI gene database isn’t limited regarding accessibility. To summarize we prefer to search for genes and organism using the NCBI gene database instead of PubMed (using text-mining) because of reliability, speed and accessibility.

Circumstances which influence anthocyanin expression

For this step we will search through biomedical literature (in PubMed). Although we mentioned above that text-mining wouldn’t be 100% accurate we will undertake several steps to drastically reduce the false positive rate. We know that the previously found genes are relatively reliable and using these while searching through PubMed will therefore reduce the amount of false positives. This will set predefined boundaries and could consequently introduce false negatives; genes which are not included in the NCBI gene database but are known in literature.

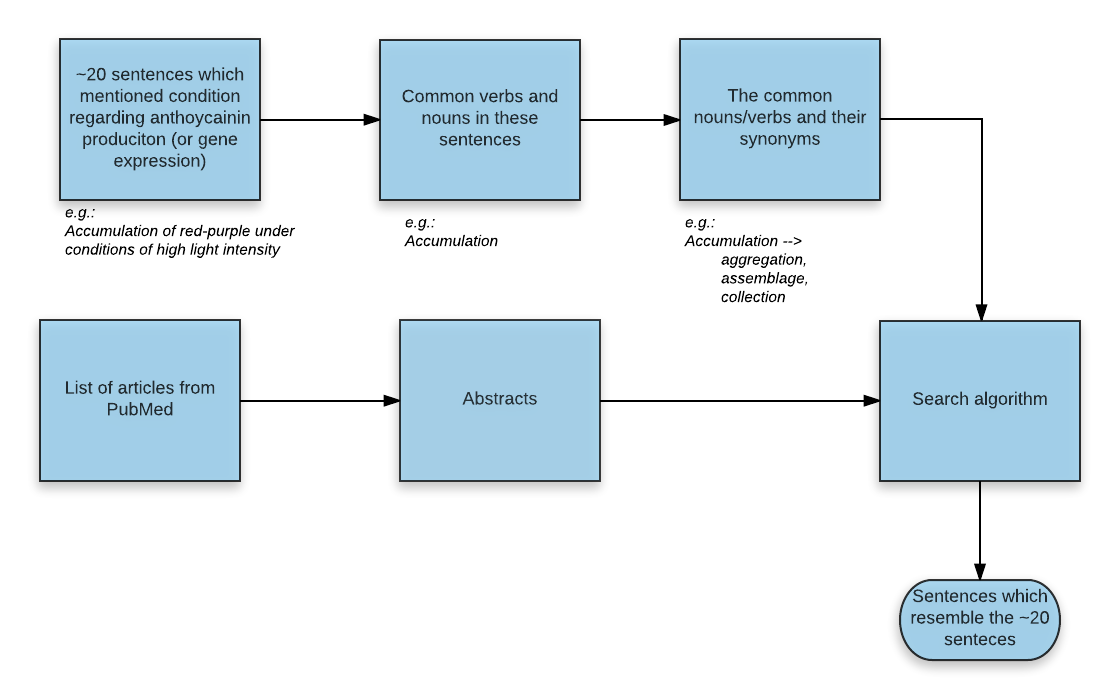
We will now explain the whole “searching” step in more detail. We will first search for the gene names + organism + anthocyanin in the PubMed database. If this doesn’t result in fulfilling matches the search will be repeated however this time without the gene names, so only organism + anthocyanin. The resulting articles are then subjected to an algorithm which retrieves the title and abstract of the articles. Another algorithm will then search for sentences which contain words indicating some sort of circumstance (e.g. condition, while, under, during, etc.) and the word “anthocyanin”(***see figure 2****)*. From these sentence the conditions will be extracted and outputted. 

Figure2, **Detailed overview of abstract analysis**. We will first “train” our algorithm to detect sentences which could mention circumstances affecting anthocyanin production. This training will be done by manually selecting approximately 20 sentences which mention circumstances from random articles. Followed by noun and verb extraction (these are the most important words in a sentence). For every verb/noun synonyms will be retrieved to extend the word list but remaining the initial meaning. These words can then be used to search through all the abstracts and select sentences which resemble the 20 sentence used during training.

Data visualization

After the retrieval of the gene names, organisms and circumstances all the data will be coupled. This will result in a graph like this:

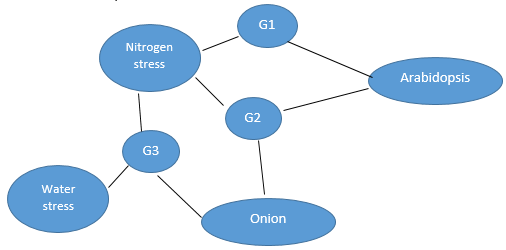


Figure3, **Expected visualization.** This figure shows the expected visualization in which all sub results are coupled into a single graph.

Table 1, **Expected Table**, This table shows the expected output (underlying the above graph)

|  |  |  |  |
| --- | --- | --- | --- |
| **Plant** | **Genes** | **Condition** | **Literature** |
| *Arabidopsis* | *G1*  *G2* | *Nitrogen stress*  *Nitrogen stress* | *Article x, article y*  *Article x* |
| *Onion* | *G3* | Nitrogen stress, Water stress | *Article z* |

***Steps***

The flowchart below shows the planned activities, each coloured area indicates activities belonging to the indicated product.

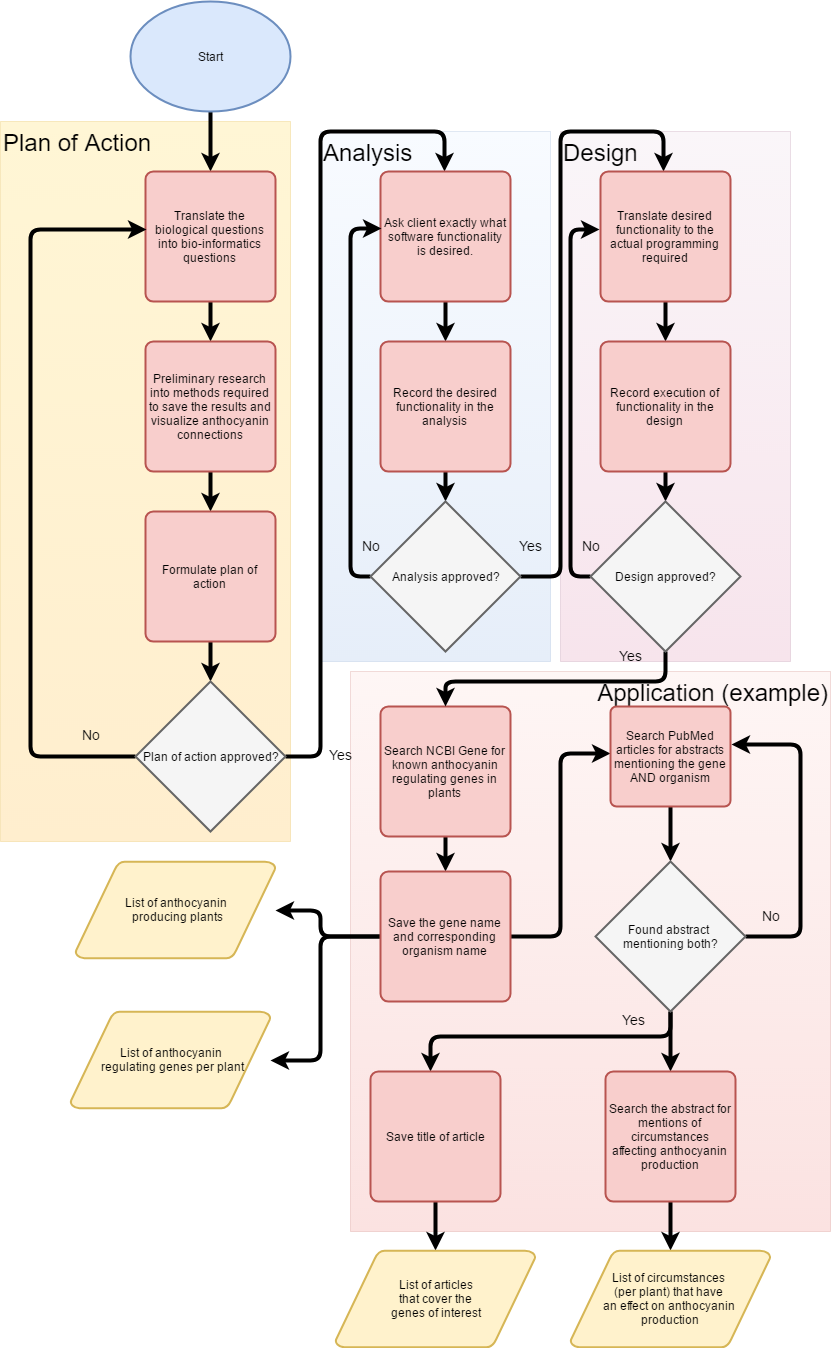


Figure 4, **Flowchart of planned activities.** This figure shows the planned activities, each coloured area indicates activities belonging to the indicated product.

The activities described in the above flowchart are also processed into a schedule:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Product** | **Wat** | **Wie** | **Hoe lang (uren)** | **Week** |
| Plan van aanpak | Inleiding | Thomas | 3 | 1 |
| Materiaal en Methode | Rick | 3 |
| Stappenplan | Koen | 3 |
| Analyse | Inleiding | Koen | 1 | 2 |
| Functionele/niet-functionele eisen | Koen | 3 |
| Use cases | Thomas | 4 |
| Architectuur | Rick | 3 |
| Ontwerp | Systeemarchitectuur | Thomas | 2 | 2-3 |
| Softwarearchitectuur | Rick | 4 |
| Technische gegevens structuur | Koen | 3 |
| Applicatie | Database | Thomas | 6 | 3-5 |
| Web interface | Koen | 8 |
| Functionaliteit | Rick | 8 |
| Poster | Inleiding | Koen | 4 | 5-6 |
| Methode | Rick | 4 |
| Resultaten | Thomas | 5 |
|  |  |  |  |  |

***References***

Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2010). Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research, 39(Database).