**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?”. We were then commissioned to develop a strategy to solve these questions.

**Approach**

To gather information about plants which are capable of anthocyanin production and to retrieve the conditions in which anthocyanin is produced (e.g. stress) two separate but closely connected steps will be conducted. First of all genes which are known to be involved in anthocyanin production will be retrieved and secondly these genes will be searched in biomedical research articles (PubMed database). The results of both steps are then combined into a table and a graph to facilitate easy interpretation. The figure below shows the different steps which will provide the answers for the following questions: “Which plants produce anthocyanin(s)?”,”which genes regulate anthocyanin production?” and “Which circumstances effect anthocyanin production?”.

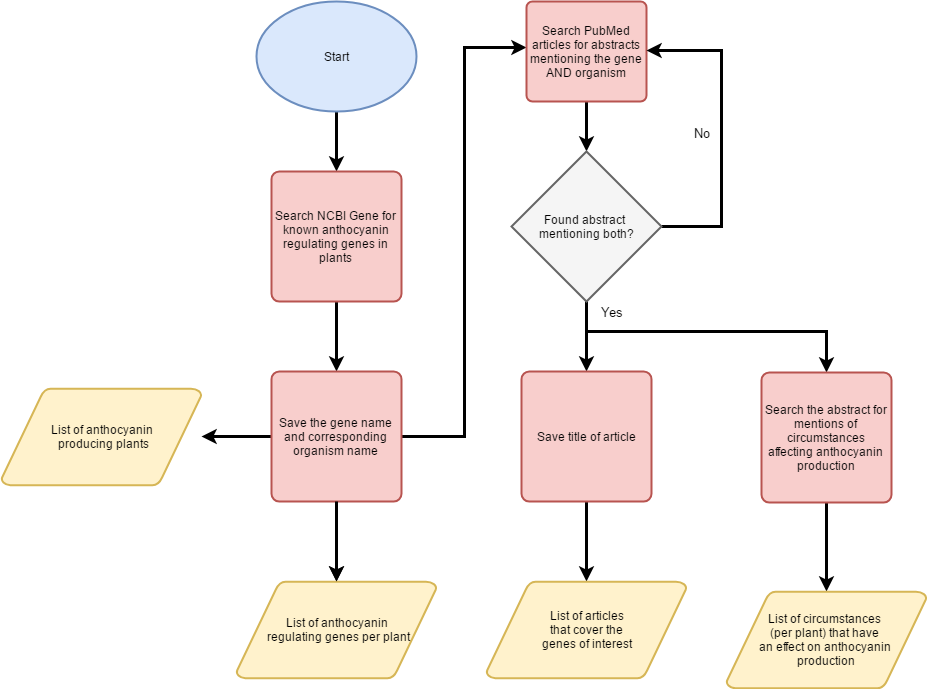


Figure1, **Flowchart**. This flowchart shows the different steps which will answer three questions: “Which plants produce anthocyanin(s)?”,”which genes regulate anthocyanin production?”, “Which circumstances effect anthocyanin production?”. A more detailed overview of “Search the abstract for mentions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

Retrieval of anthocyanin genes

The NCBI Gene 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 runtime should be considered. The NCBI gene database has predefined keywords coupled to every entry (e.g. anthocyanin in this case) making a search 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. Only the tiles, abstracts and keywords of articles can be retrieved and most of the articles are not freely available for deeper searches. On the other hand the NCBI gene search doesn’t have these limitations and allows the retrieval of all known genes. To summarize we will search the NCBI gene database for genes known to affect anthocyanin production and to infer the organism which are capable to express these.

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 reliable and using these while searching through PubMed will therefore reduce the number of false positives. This will offcourse 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 some more detail. We will first search for the gene names + organism + anthocyanin in the PubMed database. If this doesn’t result in a match the search is 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, in , while, under, during, etc.) and the word “anthocyanin”(***see figure 2****)*. From these sentence the conditions will be extracted and outputted.

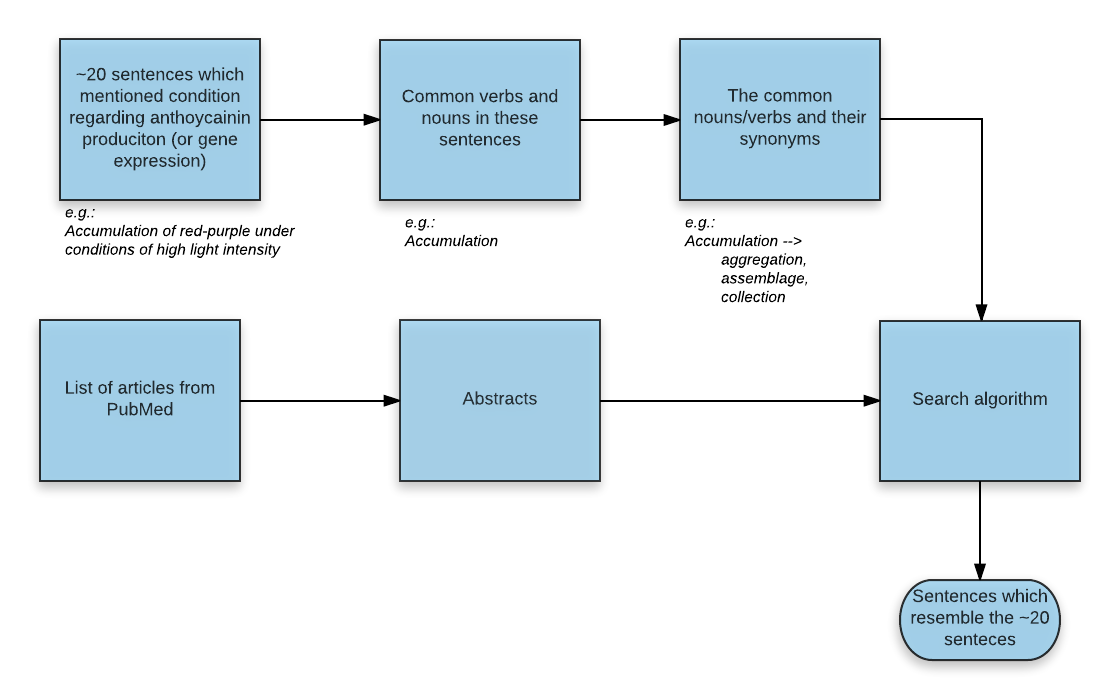


Figure2, **Detailed overview of abstract analysis**. We will first “train” our algorithm to detect sentences which could mention circumstances affecting anthocyanin production. This training will be done by manually selecting approximately 20 sentences which mention circumstances from random articles. 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:

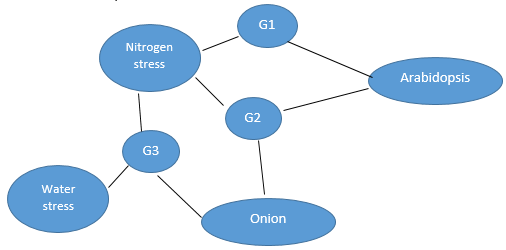


Figure3, **Expected visualization.** This figure shows the expected visualization in which all sub results are coupled into a single graph.

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* |