MSC BIF THESIS: A. THALIANA: CREATING A GENE REGULATION NETWORK IN THE SEMANTIC WEB.

Thijn van Kempen
Thijn.vankempen@wur.nl ; tlvk55@gmail.com

MSC BIF THESIS: A. THALIANA: CREATING A GENE REGULATION NETWORK IN THE SEMANTIC WEB.

Name: Thijn van Kempen (951123427020)

Course: BIF-80336

Supervisor: Harm Nijveen

Place of publication: Wageningen University, Wageningen

Date of publication: 28-2-2020

Projects GIT page: https://git.wageningenur.nl/kempe058/ara_semantics

Table of contents

Frontpage 1
Summary 3
1. Introduction 4
2. Materials and Methods 6
2.1. Retrieving of upstream promoter sequences 6
2.2. Merging the PlantPan result files 6
2.3. Filtering of PlantPan Promoter Analysis interactions 7
2.4. Deploying the semantic web 7
2.5. Tweaking the semantic web 7
2.6. Validating the semantic web 8
2.7. Storing the project and data 8
3. Results 9
3.1. Increasing minimal motif sequence length decreases number of interactions 9
3.2. Arasemantics: a new extended *A. thaliana* semantic web 10
3.3. Validation against ChIP-seq shows significant enrichment of target genes 11
4. Discussion 13
4.1. Filtering of interactions 13
4.2. Validation of the network 13
4.3. Data exploration ..... **Fout! Bladwijzer niet gedefinieerd.**
4.4. Linking Arasemantics to AraQTL 14
5. Conclusion 15
6. Acknowledgement 16
7. Glossary 17
8. References 18
9. Attachments 21

Summary

The breeding of plants with specific traits can be done by identifying the genes that influence certain variation of plant traits. Gene expression is now included as a quantitative trait for the analysis of QTLs on a genome-wide scale. Combining that knowledge with gene regulation networks might give greater insight as to how genes affect certain QTLs. However, a transcriptional gene regulation network needs to be built in order to find out which transcription factors are regulating which genes.

This network was built by scanning the promoter sequences of *A. thaliana* for transcription factor binding sites using PlantPan Promoter Analysis. Results of this analysis were filtered to reduce the amount of noise caused by hits from short binding site sequences. Filtered interactions were linked to an adaptation of the plant breeding linked data platform for *A. thaliana*. The new version of this semantic web is called Arasemantics.

After validating this network it was proven that for most transcription factors, the semantic web was significantly similar to the biological truth. A next step would be to update the transcriptional gene regulation network with data retrieved from another data source which contains information closer to the truth. After this is done, linking the semantic web to the AraQTL webtool could prove to be useful for identifying genes related to QTLs and to validate QTLS with their known master regulator.

1. Introduction

The bioinformatics and plant physiology departments of Wageningen University and Research (WUR) are interested in locating and identifying quantitative trait loci (QTL) in the genome of *Arabidopsis thaliana*. A quantitative trait is a trait that has a phenotypic variation which can be measured. This variation can be caused by environmental and/or genetic influences. These traits can be many things like the amount of leaves on a plant or the size of the fruits, but also the disease resistance of the organism (Zuo & Li, 2014) (Young, 1996). The breeding of plant species with such traits might impact the food production in a positive way. Thus making the identification of the genes belonging to such a QTL interesting to identify. The hypothesis would be that quantitative variation would be traceable to one (polymorphic) gene. The QTL mapping method has difficulties pointing out the exact location of such a gene, since there have been too few recombination events (Manly et al., 2003).

Currently, a lot of data is available on *A. thaliana,* which should make the identification of QTLs easier. Thanks to the availability of RNA sequencing and microarrays, gene expression can now be included as a quantitative trait for the analysis on QTLs on a genome-wide scale. This results in a polymorphic locus that is associated with one or more genes that have a variation in expression. This phenomenon is called "expression quantitative trait loci" (eQTL). Each of the genes that are found in the eQTL have their own eQTL profile, which is influenced by the presence of genetic variation in the regulatory elements or genes (Nijveen et al., 2017). Gene regulatory networks (GRNs) can be constructed by the usage of eQTLs (Keurentjes et al., 2007). GRNs can be used to visualize underlying relationships among genes. Those relationships can be, for example, expressional patterns or even relationships between the translated proteins.

If a genomic region influences the expression levels of (many) genes, that specific region will be called a "hotspot". Such hotspots contain points of interest when gene expressions map to the same location, since this may indicate the existence of a genetic regulator (J. Tian et al., 2016). Combining this knowledge with a tool like the TF2Network (Kulkarni et al., 2018) may improve the identification of the way genes are regulating each other in a specific eQTL. The tool for exploring eQTL patterns in *A. thaliana* is called AraQTL (Nijveen et al., 2017). With this tool, a target gene can be queried and genes that share the same expression pattern, will be shown in the output. As mentioned before, (gene)regulating networks do not only have to be based on gene expression data, they can also be made from different data such as protein-protein interactions.

The most important thing to do when building a gene regulation network is to identify the regulatory genes that are involved (Kulkarni & Vandepoele, 2019). Searching for binding site motifs in promoter sequences with the help of position weight matrices are a common approach for retrieving the interactions between regulated genes and the transcription factors. Another approach used for building gene regulation networks is to use data retrieved from a ChIP-seq experiment. The Chromatin Immunoprecipitation (ChIP) technique is a technique which is used for extracting specific protein-DNA chromatin complexes. Transcription factors can be labeled as such DNA binding proteins (Pavesi, 2017). The ChIP technique is performed on thousands of cells at the same time, to ensure that there are enough copies of the DNA regions that are bound by the transcription factor. Afterwards these locations on the genome are identified. At the moment a binding region is found on the promoter sequence, the transcription factor will most likely be a regulator of this target gene. Patterns in expression data of genes can be used to figure out if genes are regulated by a transcription factor. If there is a correlation between the expression of a transcription factor and possible targets. It could be the case that the gene expression is regulated by that transcription factor. Although ChIP-seq experiments could also result in false positives it can be seen as close to the biological truth (Pavesi, 2017). Next to ChIP-seq data there is one type of data that is even closer to the biological truth and those data are obtained from knock-out experiments. In a knock-out experiment the expression of a certain gene comes to a standstill. If the gene is responsible for the production of a transcription factor, then all the direct targets and (sometimes) unfortunately also all the indirect targets are affected as well, since they are not being regulated anymore by this particular transcription factor.

Thanks to the adapted version of pbg-ld (Tang, 2019) there is already a network tool that uses data from StringDB, KEGG and GO databases (Szklarczyk et al., 2019). This tool allows the user to input a set of genes (found on eQTL expression profile regions), so known interactions between these genes can be retrieved. The network building is done by querying a semantic web database in which the data from the previously mentioned databases is stored. As explained by Tim Berners-Lee (inventor of the World Wide Web), the semantic web is an extension of the World Wide Web, in which information is given well-defined meaning, better enabling computers and people to work in cooperation.

The goal is to expand the semantic web in such a way that links can be discovered between transcription factors and the genes that they regulate. This information combined with the expression pattern data in AraQTL could help identify which genes belong to a specific QTL. PlantPan 3.0 is a Plant Promoter Analysis Navigator tool (Chow et al., 2019), which serves as an informative resource for detecting transcription factor binding sites. The detection of the interactions between transcription and genes is the information that is needed for building the extended semantic web.

2. Materials and Methods

2.1. Retrieving of upstream promoter sequences

To make a semantic web containing links between transcription factors and the genes they regulate, more data is needed other than the data which is currently available on the semantic web (Tang, 2019). This semantic web is an adaptation of the plant breeding linked data platform called pbg-ld (Kuzniar, 2018) which is a part of the candYgene project (Visser & Finkers, 2019). To retrieve these interactions an analysis called PlantPan Promoter Analysis was run. PlantPan Promoter Analysis was chosen as a tool for retrieving transcription factor binding site information since it uses 3549 PWMs from which 1100 are validated by literature studies and the other 2449 PWMs are retrieved from ChIP-seq experiments. PlantPan Promoter Analysis uses a program called Match (Kel et al., 2003) to scan DNA sequences for potential transcription factor binding sites, Match requires PWMs to function. If a motif is detected, the hit will be documented in the PlantPan Promoter Analysis result file. Since the tool needs FASTA files that are smaller than 600kb per file, the promoter sequences FASTA file of the *A. thaliana* genome needed to be split in multiple FASTA records. Those split FASTA records where then used to perform multiple PlantPan Promoter Analysis runs. The 32833 promoter sequences of the *A. thaliana* genes were obtained from the Ensembl Plants database (Bolser et al., 2016) by downloading the 500bp upstream sequences.

2.2. Merging the PlantPan result files

Since the promoter analysis was run for every split FASTA file, multiple result files were obtained: one for each run. All these resulting files were merged into one file. This merged result file contained all the information that was needed to expand the current gene regulation network (Tang, 2019) and add the transcriptional regulatory network. The merged PlantPan result file contains information on the promoter sequence in which the transcription factor binding site was detected. The transcription factor matrix ID, as well as the motif sequence that was used for the detection of the transcription factor were available. Furthermore, it shows the transcription factor ID and the transcription factor family belonging to that ID. The columns transcription factor ID and the promoter sequence ID were required for building the turtle file, and thus for forms the basis for the gene regulation network.

2.3. Filtering of PlantPan Promoter Analysis interactions

After retrieving the PlantPan Promoter Analysis results they needed to be filtered. Since transcription factor binding sites are on average 10 nucleotides long, the filtering was applied to short binding site motif sequences (Stewart & Plotkin, 2012). A lot of the hits that were found during the analysis contained hits on motif binding sites with a motif sequence that was shorter than 7 bases. The hits based on these short binding site motifs were all removed from the result file since they were seen as possible noise for the gene regulation network. Storing all unfiltered results in the semantic web would make the data retrieval unnecessary slow compared to the time it takes when these hits are removed.

To make sure that the shorter motifs were causing noise in the dataset, and removing them was the right choice, the number of times a motif sequence was found in the promoter sequences was investigated. Theoretically speaking, when a motif sequence is 5 bases long, a random hit in the promotor sequences occurs once every 1024 bases.

Since there are 32833 promoter sequences with a length of 500 bases each, the motifs should, theoretically speaking, be found around 16030 times. This filtering step was done to see if the amount of times a motif has been found differs from the expected times a motif is found. If the number does not match, it means that the motif sequence is found less/more times than expected.

2.4. Deploying the semantic web

First a Docker container (*Docker Container*, 2020) was created by deploying the already existing semantic web (Tang, 2019). This Docker image allows the user to build a nonrestrictive environment for the semantic web. The data virtualization platform Virtuoso (*Virtuoso*, 2018), which is run from within the Docker container, stores the semantic web. This specific file format is used for storing RDF data, therefore the PlantPan Promoter Analysis result file was converted to the turtle file format.

2.5. Tweaking the semantic web

After running the PlantPan Promoter Analysis, many interactions were found between transcription factors and potential targets. Due to querying restrictions on the semantic web, the database required some optimization. The optimization mainly occurred on the database management platform, allowing the user to perform more CPU intensive queries. This was achieved by raising the maximum memory usage, increasing the query execution and timeout limit time. Furthermore queries that returned more than 1000 results can now be returned completely due to an increase in the maximum row result set. The maximum row result set was increased to 50000 since there are transcription factors that regulate many target genes.

2.6. Validating the semantic web

ChIP-seq datasets were retrieved for the transcription factors ABI5, PLT2, SEP3 and WIND1. These ChIP-seq datasets contained information on which target each transcription factor has according to the ChIP-seq peaks. These targets were then compared to targets obtained from the PlantPan analysis, to find out if the targets were similar. This was done by calculating the Jaccard Index between the dataset, to obtain the similarity scoring and by running a Fisher exact test on the datasets. The Fisher exact test was performed to see whether the similar interactions between the two datasets were significant. Furthermore, targets retrieved by AraQTL for transcription factor ABI5 were compared in the same way to the targets retrieved by PlantPan. Finally, both the AraQTL and the PlantPan target dataset were used as an input to produce a network with TF2Network. This was done as a validation to figure out how many targets were targets according to the prediction of a different tool. As well as PlantPan, TF2Network scans promoter sequences with PWM of motif sequences to detect transcription factor binding sites.

2.7. Storing the project and data

A GitHub page (van Kempen, 2020) with all the scripts and data which were needed to perform this extension of the gene regulation network was made. Furthermore additional scripts and data were added to this GitHub page which were then used to either produce graphical presentation of/or used for filtering the PlantPan result data. The new extended semantic web is still run on a Docker image, which simplifies the deployment of the semantic web. It can be run on a server or private computer.

3. Results

3.1. Increasing minimal motif sequence length decreases number of interactions

After running the PlantPan Promoter Analysis, 71.5 million interactions were found between transcription factors and targets. It is said that transcription factors account for 9% of *A. thaliana* protein encoding genes, since there are 27655 protein coding genes, of which around 2500 account for transcription factors (Van Leene et al., 2016). Literature has shown that *A. thaliana* transcription factors typically regulate between 417 and 6028 direct targets (Brooks et al., 2019). This indicates that by using these numbers, the amount of interactions found by the PlantPan Promoter Analysis should be between one million and 15.1 million. The number of interactions found by the PlantPan analysis was around 71.5 million, this indicates that there are a lot more predicted interactions. Filtering was applied to the PlantPan result set. This filtered the short motif binding site sequences as described in section 2.3. The network is based on interactions with a minimal motif length of 7 bases. This filtering step reduced the number of found interactions from 71.5 million to 20.2 million. This reduced the data by 71.4% as shown in figure 1. The largest reduction in interactions seems to be when the filter applied a minimal motif length of 6bp. This reduced the dataset from 71.5 million to 23.5 million interactions. It appears that a lot of interactions between transcription factors and targets were based on hits of motif binding site sequences that were equal to 5 bases.

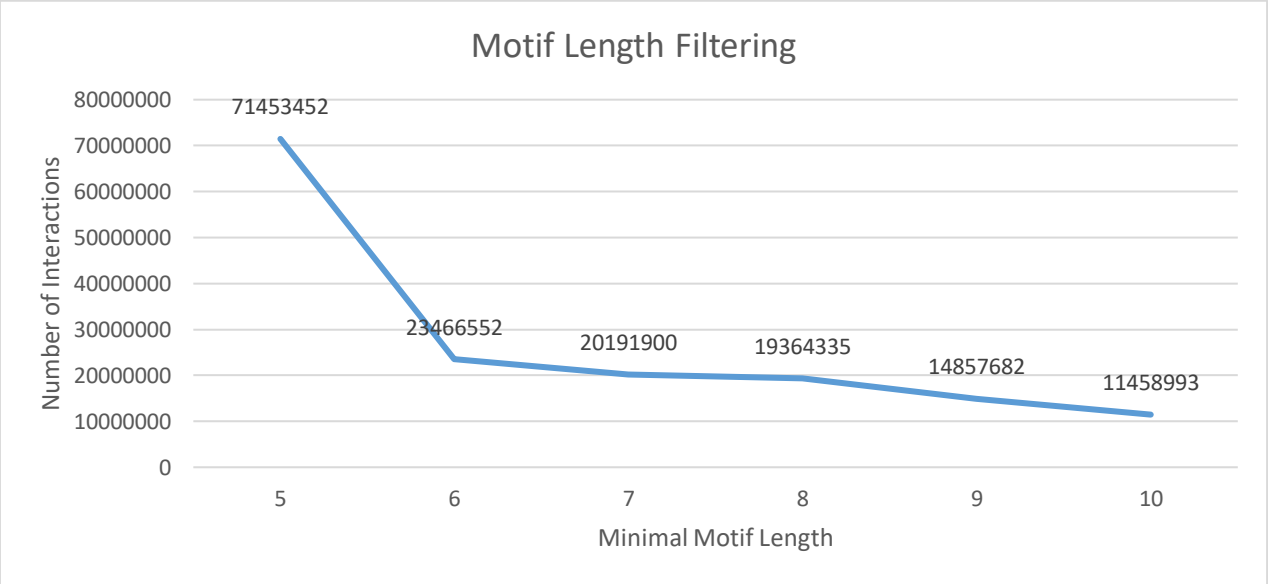

Figure 1 The reduction of the interaction counts of the PlantPan Promoter Analysis result file, based on minimal motif length filtering. This distribution plot shows the distribution of the interactions among minimal motif lengths.

The reduction of this filtering filtered out a vast amount of target hits. Figure 3 shows the number of times a motif was found in the promoter sequences. In figure 3a, the largest outlier of the group was found more than 1.75 million times. As shown in section 2.3, the expectation of the number of times a motif would be found with the length of 5 bases would be around 16030 times. However, this particular motif was found 109 times more often than expected. While the largest outlier in figure 3b was discovered around 80000 times, which is found less than 5 times as often. This could indicate that increasing the minimal motif length from 5 to 7 bases causes a lot of faulty interactions to disappear. Figure 4 shows the amount of targets each transcription factor has, 4b shows that even after filtering there are more interactions found then the average of literature studies suggest (Brooks et al., 2019).

3.2. Arasemantics: a new extended *A. thaliana* semantic web

After retrieving and filtering the data that was needed for building a transcription factor regulatory network, the network could be extended. The new transcriptional regulatory interactions were linked to the data that already existed in the semantic web (Tang, 2019). This allows the user to now not only explore the *A. thaliana* protein-protein interaction network, but also explore the *A. thaliana* transcription factor regulatory network in the already existing semantic web. This new version of the extended semantic web is called Arasemantics (van Kempen, 2020).

The semantic web can be explored by querying it through a web interface which allows the user to explore the data records. An example of such a record could be a record for a gene ID as shown in figure 2. This record contains information such as the location of the gene and a description belonging to the gene. When the gene is a transcription factor which regulates other genes, a separate category becomes available. The category “001154”, contains the targets predicted by the PlantPan Promoter Analysis for the specific transcription factor. Category 001154 was an already existing regulation category, found in the Semanticscience Integrated Ontology (SIO) (Dumontier et al., 2014), which is an ontology that facilitates biomedical knowledge discovery. SIO offers relations and classes which describe objects, processes and their attributes in the biomedical domain.

Attributes	Values
type	protein_coding_gene http://rdf.ebi.ac.uk/terms/ensembl/protein_coding
label	ABI5
seeAlso	http://identifiers.org/ensembl/AT2G36270
dc:description	Protein ABSCISIC ACID-INSENSITIVE 5 [Source:UniProtKB/Swiss-Prot;Acc:Q9SJN0]
dc:identifier	AT2G36270
location	chromosome 2:15204659-15207636:-1
http://semanticsci...g/resource/001154	http://rdf.ebi.ac.uk/resource/ensembl/AT3G63040 PYL9 GSTF6 http://rdf.ebi.ac.uk/resource/ensembl/AT1G03520 ERF094 »more»

Figure 2 A record for gene ID ABI5 obtained by querying the semantic web. It shows all the (linked) information available on ABI5.

3.3. Validation against ChIP-seq shows significant enrichment of target genes

Test cases were made in which certain transcription factors (ABI5, PLT2, SEP3, WIND1) and the genes that they regulate were retrieved from the semantic web and then compared to existing ChIP-seq datasets (Bi et al., 2017; Iwase et al., 2011; Kaufmann et al., 2009; Santuari et al., 2016). The comparison is done to ChIP-seq data since ChIP-seq experiments only mark targets as actual targets when reads that belong to binding sites are significantly enriched (Pavesi, 2017). Overlap between the targets from the PlantPan Promoter Analysis and the targets from the ChIP-datasets should be expected since the PlantPan tool uses PWMs obtained from ChIP-seq experiments (Chow et al., 2019). The transcription factors ABI5, PLT2, SEP3 and WIND1 are well studied and are known to be regulators of many genes. ABI5 is a transcription factor that plays a key role in the regulation of seed germination (Skubacz et al., 2016). PLT2 plays a role in guiding the progression of cell differentiation in different parts of the *A. thaliana* root (Santuari et al., 2016). SEP3 binds to thousands of targets which play a role in hormonal and growth related pathways (Kaufmann et al., 2009). WIND1 is involved in cell dedifferentiation (Iwase et al., 2011).

This comparison gave an overview of the interactions predicted by the PlantPan Promoter Analysis and if they matched the validated interactions retrieved from ChIP-seq experiments. This validation was done to test whether the PlantPan Promoter Analysis was a suitable one to base a transcription factor to target regulation network on. Measuring the similarity between the two target sets was done by calculating the Jaccard Index (Levandowsky & Winter, 1971). Furthermore, the predictions from PlantPan were compared to the genes that shared the same expression patterns as ABI5 which were predicted by the AraQTL webtool for the transcription factor ABI5. The transcription factor ABI5 (Skubacz et al., 2016) was chosen, since this is a well-studied transcription factor which is responsible for the regulation of multiple genes. Both the AraQTL genes of interest and the ChIP-seq targets which were retrieved from different papers were compared to the targets of the transcription factors retrieved by the PlantPan prediction. A Fisher exact test was performed to see whether the overlapping targets between the datasets were found to be significant.

A query was performed on different transcription factors to retrieve their targets, which were predicted by the PlantPan Promoter Analysis. The transcription factors that were investigated are: ABI5, PLT2, SEP3 and WIND1. These targets were compared to a list of ChIP-seq targets datasets (Bi et al., 2017; Iwase et al., 2011; Kaufmann et al., 2009; Santuari et al., 2016), which were obtained from literature studies.

Evaluation was done on the overlap of the semantic web interactions (obtained from the PlantPan Promoter Analysis) and ChIP-seq datasets. This overlap is used as an indication for the quality of the interactions that were obtained with the PlantPan Promoter Analysis. Table 1 presents an overview of information, obtained by calculating the Jaccard Index and by performing a Fisher exact test on the two data sets. The Jaccard Index is calculated by dividing the intersection with the union of the datasets. Its value is the ratio of similarity between the two datasets. The Fisher exact test was performed on the tables 3 to 6 to determine if the similar targets are found to be significant. As shown in table 1 the Jaccard Indexes are rated from 0.0013 to 0.0995, indicating that the overlap between all the datasets of the different transcription factors is all small. While the Jaccard Index is low, the P-values retrieved by performing a Fisher exact test are found to be significant between the targets of transcription factor ABI5, SEP3 and WIND1. Indicating that the similar interactions found between transcription factor and targets in the two datasets are found to be significantly enriched.

Table 1 The comparison of transcription factor targets obtained by querying the semantic web and comparing the obtained data with ChIP-seq targets. The table shows information on the intersection and union between the two datasets and shows the calculated Jaccard Index. Results marked with a "" were obtained by performing a Fisher exact test on the two datasets.*

Transcription factor	ABI5	PLT2	SEP3	WIND1
Semantic web targets	3919	240	11122	1407
ChIP-seq targets	16	355	3475	2399
Intersection	5	3	1321	126
Union	3930	533	13276	3680
Jaccard Index	0.0013	0.0056	0.0995	0.0342
P-Value *	0.0341	0.7472	8.68e-13	0.0157

Possible targets of transcription factor ABI5 were retrieved with AraQTL. The targets of ABI5 were retrieved by selecting the genes that shared similar expression patterns as ABI5 with a correlation threshold of 0.85 and the experiment set Joosen_etal_2012 (Joosen et al., 2012). This resulted in a set of 304 AraQTL targets, which were compared to the targets from the semantic web. This comparison was done to find out if the semantic web could be a way to find new or validate previously found master regulators for eQTL patterns. After calculating a Jaccard Index and performing a Fisher exact test on the two datasets. It was observed in table 2 that the dataset have a low Jaccard Index and a very low P-value. This shows that the targets between the two datasets are found to be significantly enriched.

Table 2 The comparison of ABI5 targets obtained by querying the semantic web and comparing the obtained data with AraQTL targets. The table shows information on the intersection and union between the two datasets and shows the calculated Jaccard Index. Results marked with a "" were obtained by performing a Fisher exact test on the two datasets.*

Transcription factor	ABI5
Semantic web targets	3919
AraQTL targets	304
Intersection	115
Union	4108
Jaccard Index	0.028
P-Value *	1.66e-31

The targets obtained from AraQTL were used as input data for creating a network with TF2Network. This resulted in the creation of a transcription factor to target network with ABI5 as the transcription factor as shown in figure 6. As shown in table 2 there were 304 AraQTL targets, from those targets 158 were used for the creation of the ABI5 gene regulation network. This indicates that over half the potential targets retrieved by AraQTL are recognized, according to transcription factor databases, as actual regulators. Figure 7 shows that 99 out of the 304 AraQTL genes were being targeted by a specific motif sequence. This suggests that a lot of genes that play a role in seed germination share a similar motif binding sequence.

Another network for ABI 5 was created with TF2Network with the potential 3919 targets, which were predicted by PlantPan as input data. This new network contained 2354 targets out of the 3919 targets from the PlantPan dataset. This indicates that for this particular test case around 60% of the predicted targets were recognized by the prediction algorithm of TF2Network.

4. Discussion

4.1. Filtering of interactions

The interactions found after applying the filter step on the minimal motif sequence length of 7 bases were reduced by 71.4% as shown in section 3.1. Although a lot of interactions with shorter, less specific motif sequences were removed It is still a question whether this was the best filtering step for the input data of the semantic web. An option would be to dig into the transcription factors that regulate above ten thousand targets. If the Plant Pan Promoter Analysis retrieves transcription factors that regulate a huge amount of genes it is debatable if the interactions are trustworthy and also if such interactions add any value to the network. Filtering can also occur on motifs sequences that are identified in a very large proportion of the promoter sequences. If a motif sequence is found multiple times, it could suggest that it is just a repetitive sequence found among many genes which, if not filtered out, would introduce a lot of noise in the semantic web.

4.2. Validation of the network

By calculating a Jaccard Index it was shown that the Jaccard Index is not a good way of testing if your network is similar to the biological ChIP-seq representation. This is mostly due to the large differences in size of the dataset, as shown in table 1. The Jaccard Index is calculated by dividing the intersection by the union. If the intersection is low in comparison to the union and this is due to size differences in the two datasets, you automatically end up with a low Jaccard index. This indicates a low similarity between the sets. However, a Fisher exact test works in a different way and is not affected by different sample sizes. It calculates whether the interactions that are similar in both datasets are significant or not. This Fisher exact test gave a good insight on whether the semantic web was found to be significantly similar to a biological representation of transcription factor regulatory networks.

Table 1 shows that three out of the four investigated transcription factors have similar targets between the semantic web and ChIP-seq datasets that are found to be significant. Showing that part of the interactions between transcription factors and targets of the semantic web are significantly enriched among the ChIP-seq datasets. However, the huge differences in sample sizes among the datasets (see table 1) show that there are a lot of interactions that are not shared between the ChIP-seq datasets and the semantic web. Comparing different transcription factors and their targets which are obtained with the AraQTL webtool and the semantic web might result in different outcomes.

The possible targets that share the same expression pattern as ABI5 which were found with AraQTL are significantly enriched when they are compared to the targets of the semantic web. This could indicate that the targets obtained from the PlantPan Promoter Analysis could overall serve as a good representation of the ABI5 regulation network. These results might not give a proper biological true representation of the transcription factor regulatory network, since there are more interactions found with the PlantPan Promoter Analysis. This indicates that a different method for retrieving the data needed to create the network may need to be developed. A tool that also mentions the statistical power of interactions would give greater insight as to whether the interaction could be true or false. Such a tool could be plantregmap (F. Tian et al., 2020). This tool functions for most part the same as the PlantPan Promoter Analysis. Both tools require promoter sequences as input and scan for transcription factor binding sites with motifs. With plantregmap however, the user can provide a p-value threshold, which would hopefully obtain more truthful interactions. Rebuilding the network with better interactions should give the semantic web even more power. Although the ABI5 test case has proven that most of the targets found in the semantic web are enriched, there is still no golden test set available in which you can truly validate the network perfectly. For now it is proven that the semantic web can be used to retrieve the targets of multiple transcription factors.

4.3. Exploring the semantic web

At this moment the exploration of the semantic web is done by querying the semantic web. However, not everyone who is interested in exploring the gene regulatory network of *A. thaliana* is familiar with SPARQL (Tally, 2010) and the structure of the data. This shows that Arasemantics is not user friendly at the moment. Therefore a better data exploration approach should be developed. The best option would be to make a graphical user interface in which the user can upload a set of genes and in which a result gets displayed whilst showing all the interactions for those particular genes as well as information on those genes. A way to display such a network could be in the same way the TF2Network (Kulkarni et al., 2018) visualizes their interactions as shown in figure 5. Combining the information that is found within a gene ID record (see figure 2), with nodes of the network with the interactions of those particular genes would be a good first step towards user friendliness.

4.4. Linking Arasemantics to AraQTL

Currently AraQTL can provide the user with a set of genes that share the same expression pattern as the gene that you are interested in. Querying this list of genes to see which genes are connected and how they are interacting with themselves and other genes might give a better insight as to how the transcriptional regulatory network is connected for different eQTLs. Connecting the AraQTL web interface with a new graphical user interface for Arasemantics could prove useful for both validating the transcriptional regulatory network and exploring the data stored in the semantic web.

5. Conclusion

The sematic web is expanded with a transcriptional regulatory network and is now called Arasemantics. This transcriptional regulatory network is filled with interactions that are obtained from the PlantPan Promoter Analysis. It is difficult to see how good the network represents the biological truth since some interactions have not been validated yet and some of the genes and transcription factors are currently not well known. After running multiple validations it was shown that the targets from three out of four transcription factors were significantly enriched. This result gave an indication that Arasemantic could function as a tool for studying the transcriptional regulatory network for most *A. thaliana* transcription factors. However, the recollecting of data for the transcriptional regulatory networks might prove useful to obtain a network that is closer to the biological truth. It also important to keep the interactions in Arasemantics up to date since there is more information publicly available on *A.thaliana* every day. Furthermore, linking the AraQTL webtool and Arasemantics together should make the exploration of identifying genes that belong to an eQTL easier.

6. Acknowledgement

I would like to give my thanks towards the students of the Bioinformatics department who were always ready to go into discussion about the project and provided their opinions on the matter. Also thanks to some of the students who were willing to give me feedback during the preparation of the midterm presentation. Off course I would also like to thank the group for the great informal discussions. There are three people that I would like to thank especially, namely Weiqi Tang and Harm Nijveen and Aalt-Jan van Dijk.

Weiqi Tang is the student who adapted the original plant breeding linked data platform and who created the linked data platform for *A. thaliana* (Tang, 2019). Furthermore, Weiqi spend some time during and even after finishing his project to help get me into linked data and get me started with my first SPARQL queries.

Harm Nijveen was my supervisor during this project with whom I had great discussions related and unrelated to this project. He was the person that got me interested in QTLs and who came with the idea to make a linked data platform for *A. thaliana*. Furthermore Harm is also the creator of AraQTL (Nijveen et al., 2017). Harm guided me through the project and always showed interest in the ideas that were brought to the table even though they were not always that significant. Harm gave me the space and trust that I needed to really make this project my own.

Aalt-Jan van Dijk gave me ideas and talked to me about validating the input data for my semantic web. This gave me the idea how to use the Fisher exact test on the different datasets to see whether my targets were significantly enriched among other test cases.

7. Glossary

ChIP-seq: ChIP-sequencing, a method used for analyzing protein interactions within DNA.

Docker: A software tool for virtualizing software in packages.

Pbg-ld: Plant breeding and genomics linked data platform.

Quantitative Trait Loci (QTL): A locus which correlates with a quantitative trait that causes variation in the phenotype in a population of a certain organism.

Resource Description Framework (RDF): A family of the World Wide Web consortium which serves as a metadata data model. In this framework relationships between data objects are described which give meaning to the data object.

Semantic web: Extension of the World Wide Web, it enables the encodings of semantics through the RDF and Web Ontology Language (OWL).

SPARQL: A query language used for querying a semantic web.

Turtle: The file format used for expressing data in the RDF model.

Virtuoso: A data virtualization platform for storing data in a structured data model.

8. References

Bi, C., Ma, Y., Wu, Z., Yu, Y. T., Liang, S., Lu, K., & Wang, X. F. (2017). Arabidopsis ABI5 plays a role in regulating ROS homeostasis by activating CATALASE 1 transcription in seed germination. *Plant Molecular Biology*, *94*(1–2), 197–213. https://doi.org/10.1007/s11103-017-0603-y

Bolser, D., Staines, M. D., Pritchard, E., & Kersey, P. (2016). *Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomics Data.* https://link.springer.com/protocol/10.1007%2F978-1-4939-3167-5_6

Brooks, M. D., Cirrone, J., Pasquino, A. V., Alvarez, J. M., Swift, J., Mittal, S., Juang, C. L., Varala, K., Gutiérrez, R. A., Krouk, G., Shasha, D., & Coruzzi, G. M. (2019). Network Walking charts transcriptional dynamics of nitrogen signaling by integrating validated and predicted genome-wide interactions. *Nature Communications*, *10*(1), 1–13. https://doi.org/10.1038/s41467-019-09522-1

Chow, C. N., Lee, T. Y., Hung, Y. C., Li, G. Z., Tseng, K. C., Liu, Y. H., Kuo, P. L., Zheng, H. Q., & Chang, W. C. (2019). Plantpan3.0: A new and updated resource for reconstructing transcriptional regulatory networks from chip-seq experiments in plants. *Nucleic Acids Research*, *47*(D1), D1155–D1163. https://doi.org/10.1093/nar/gky1081

Docker Container. (2020). https://www.docker.com/resources/what-container

Dumontier, M., Baker, C. J. O., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., Del Rio, N. R., Duck, G., Furlong, L. I., Keath, N., Klassen, D., McCusker, J. P., Queralt-Rosinach, N., Samwald, M., Villanueva-Rosales, N., Wilkinson, M. D., & Hoehndorf, R. (2014). The semanticscience integrated ontology (SIO) for biomedical research and knowledge discovery. *Journal of Biomedical Semantics*, *5*(1), 1–11. https://doi.org/10.1186/2041-1480-5-14

Iwase, A., Mitsuda, N., Koyama, T., Hiratsu, K., Kojima, M., Arai, T., Inoue, Y., Seki, M., Sakakibara, H., Sugimoto, K., & Ohme-Takagi, M. (2011). The AP2/ERF transcription factor WIND1 controls cell dedifferentiation in arabidopsis. *Current Biology*, *21*(6), 508–514. https://doi.org/10.1016/j.cub.2011.02.020

Joosen, R. V. L., Arends, D., Willems, L. A. J., Ligterink, W., Jansen, R. C., & Hilhorst, H. W. M. (2012). Visualizing the genetic landscape of Arabidopsis seed performance. *Plant Physiology*, *158*(2), 570–589. https://doi.org/10.1104/pp.111.186676

Kaufmann, K., Muiño, J. M., Jauregui, R., Airoldi, C. A., Smaczniak, C., Krajewski, P., & Angenent, G. C. (2009). Target genes of the MADS transcription factor sepallata3: Integration of developmental and hormonal pathways in the arabidopsis flower. *PLoS Biology*, *7*(4), 0854–0875. https://doi.org/10.1371/journal.pbio.1000090

Kel, A. E., Gößling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., & Wingender, E. (2003). MATCH^{TM}: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research*, *31*(13), 3576–3579. https://doi.org/10.1093/nar/gkg585

Keurentjes, J. J. B., Fu, J., Terpstra, I. R., Garcia, J. M., Van Den Ackerveken, G., Snoek, L. B., Peeters, A. J. M., Vreugdenhil, D., Koornneef, M., & Jansen, R. C. (2007). Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(5), 1708–1713. https://doi.org/10.1073/pnas.0610429104

Kulkarni, S. R., & Vandepoele, K. (2019). Inference of plant gene regulatory networks using data-driven methods: A practical overview. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 194447. https://doi.org/10.1016/j.bbagrm.2019.194447

Kulkarni, S. R., Vaneechoutte, D., Van de Velde, J., & Vandepoele, K. (2018). TF2Network: predicting transcription factor regulators and gene regulatory networks in Arabidopsis using publicly available binding site information. *Nucleic Acids Research*, *46*(6), e31–e31. https://doi.org/10.1093/nar/gkx1279

Kuzniar, A. (2018). *pbg-ld*. https://zenodo.org/record/1458169

Levandowsky, M., & Winter, D. (1971). Distance between sets [5]. *Nature*, *234*(5323), 34–35. https://doi.org/10.1038/234034a0

Manly, K. F., Farber, C. R., Shou, S. M., Van Zant, G., Bachmanov, A. A., Nowakowski, R. S., Spearow, J. L., Gu, W. K., Blankenhorn, E. P., Blizard, D. A., Paigen, B., Hunter, K., Threadgill, D. W., Gershenfeld, H., Iraqi, F. A., de Haan, G., Montagutelli, X., Bureau, J. F., Doerge, R. W., … Chesler, E. J. (2003). The nature and identification of quantitative trait loci: a community's view. *Nature Reviews Genetics*, *4*(11), 911–916.

Nijveen, H., Ligterink, W., Keurentjes, J. J. B., Loudet, O., Long, J., Sterken, M. G., Prins, P., Hilhorst, H. W., de Ridder, D., Kammenga, J. E., & Snoek, B. L. (2017). AraQTL – workbench and archive for systems genetics in Arabidopsis thaliana. *Plant Journal*, *89*(6), 1225–1235. https://doi.org/10.1111/tpj.13457

Pavesi, G. (2017). ChIP-Seq data analysis to define transcriptional regulatory networks. In *Advances in Biochemical Engineering/Biotechnology* (Vol. 160). https://doi.org/10.1007/10_2016_43

Santuari, L., Sanchez-Perez, G. F., Luijten, M., Rutjens, B., Terpstra, I., Berke, L., Gorte, M., Prasad, K., Bao, D., Timmermans-Hereijgers, J. L. P. M., Maeo, K., Nakamura, K., Shimotohno, A., Pencik, A., Novak, O., Ljung, K., van Heesch, S., de Bruijn, E., Cuppen, E., … Heidstra, R. (2016). The PLETHORA gene regulatory network guides growth and cell differentiation in Arabidopsis roots. *Plant Cell*, *28*(12), 2937–2951. https://doi.org/10.1105/tpc.16.00656

Skubacz, A., Daszkowska-Golec, A., & Szarejko, I. (2016). The role and regulation of ABI5 (ABA-insensitive 5) in plant development, abiotic stress responses and phytohormone crosstalk. *Frontiers in Plant Science*, *7*(DECEMBER2016), 1–17. https://doi.org/10.3389/fpls.2016.01884

Stewart, A. J., & Plotkin, J. B. (2012). Why transcription factor binding sites are ten nucleotides long. *Genetics*, *192*(3), 973–985. https://doi.org/10.1534/genetics.112.143370

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., & Von Mering, C. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, *47*(D1), D607–D613. https://doi.org/10.1093/nar/gky1131

Tally, R. J. (2010). Learning SPARQL. In *CLCWeb - Comparative Literature and Culture* (Vol. 12, Issue 1). https://doi.org/10.7771/1481-4374.1572

Tang, W. (2019). *pbg-ld arabidopsis*. https://github.com/Will0will/pbg-ld

Tian, F., Yang, D. C., Meng, Y. Q., Jin, J., & Gao, G. (2020). PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Research*, *48*(D1), D1104–D1113. https://doi.org/10.1093/nar/gkz1020

Tian, J., Keller, M. P., Broman, A. T., Kendziorski, C., Yandell, B. S., Attie, A. D., & Broman, K. W. (2016). The

dissection of expression quantitative trait locus hotspots. *Genetics*, *202*(4), 1563–1574. https://doi.org/10.1534/genetics.115.183624

van Kempen, T. (2020). *Arasemantics*. https://git.wageningenur.nl/kempe058/ara_semantics

Van Leene, J., Blomme, J., Kulkarni, S. R., Cannoot, B., De Winne, N., Eeckhout, D., Persiau, G., Van De Slijke, E., Vercruysse, L., Vanden Bossche, R., Heyndrickx, K. S., Vanneste, S., Goossens, A., Gevaert, K., Vandepoele, K., Gonzalez, N., Inzé, D., & De Jaeger, G. (2016). Functional characterization of the Arabidopsis transcription factor bZIP29 reveals its role in leaf and root development. *Journal of Experimental Botany*, *67*(19), 5825–5840. https://doi.org/10.1093/jxb/erw347

Virtuoso. (2018). https://virtuoso.openlinksw.com/

Visser, R. G. F., & Finkers, R. (2019). *CANDYGENE*. https://github.com/candYgene

Young, N. D. (1996). Qtl Mapping and Quantitative. *Annual Review Phytopathol*, *34*, 479–501.

Zuo, J., & Li, J. (2014). Molecular Genetic Dissection of Quantitative Trait Loci Regulating Rice Grain Size. *Annual Review of Genetics*, *48*(1), 99–118. https://doi.org/10.1146/annurev-genet-120213-092138

9. Attachments

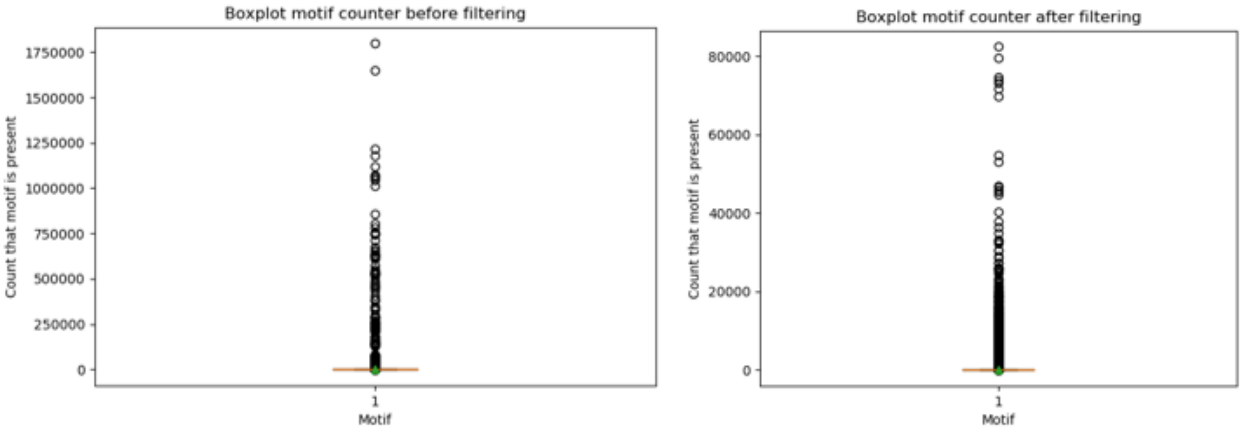

Figure 3a/b. A boxplot showing the distribution of how many times each motif is detected among the A.thaliana *promoter sequences. Each dot represents a unique motif. (3a) The data used contains all the raw nonfiltered interactions found by PlantPan. (3b) The data used contains all the filtered interactions found by PlantPan with a minimal motif length of 7bp.*

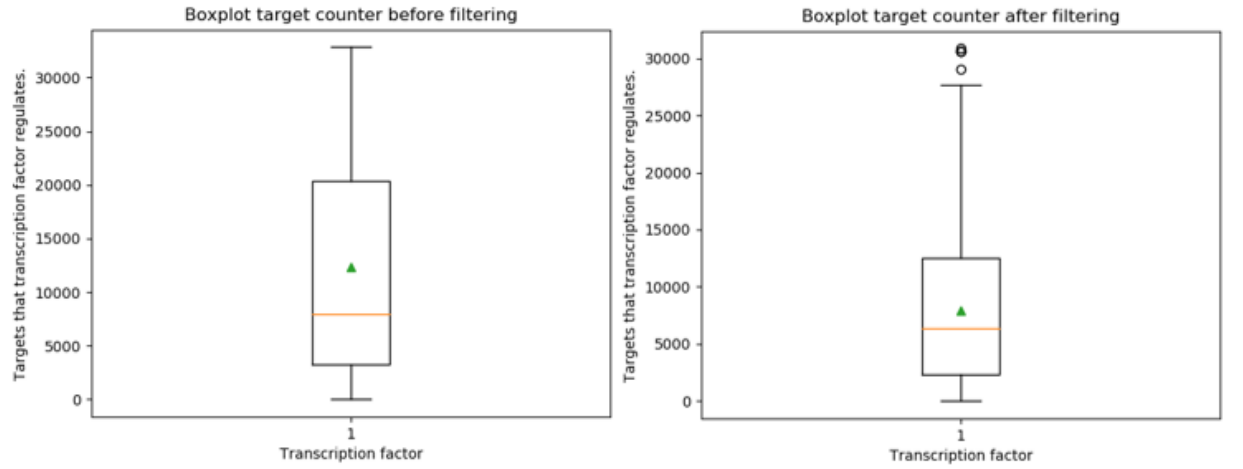

Figure 4a/b A boxplot showing the distribution of how many times targets each transcription factor has. Each dot represents the target count for an individual transcription factor. (4a) The data used contains all the raw nonfiltered interactions found by PlantPan. (4b) The data used contains all the filtered interactions found by PlantPan with a minimal motif length of 7bp.

Table 3 The matrix that was built for performing a Fisher exact test. The data in the table shows a comparison of the targets of transcription factor ABI5 retrieved from the Semantic Web (PlantPan interactions) against targets validated by ChIP-seq experiments.

ABI5	**In PlantPan**	**Not in PlantPan**
In ChIP-seq	**5**	**11**
Not in ChIP-seq	**3914**	**28908**

Table 4 The matrix that was built for performing a Fisher exact test. The data in the table shows a comparison of the targets of transcription factor PLT2 retrieved from the Semantic Web (PlantPan interactions) against targets validated by ChIP-seq experiments.

PLT2	In PlantPan	Not in PlantPan
In ChIP-seq	3	352
Not in ChIP-seq	237	32244

Table 5 The matrix that was built for performing a Fisher exact test. The data in the table shows a comparison of the targets of transcription factor SEP3 retrieved from the Semantic Web (PlantPan interactions) against targets validated by ChIP-seq experiments.

SEP3	In PlantPan	Not in PlantPan
In ChIP-seq	1321	2154
Not in ChIP-seq	9801	20878

Table 6 The matrix that was built for performing a Fisher exact test. The data in the table shows a comparison of the targets of transcription factor WIND1 retrieved from the Semantic Web (PlantPan interactions) against targets validated by ChIP-seq experiments.

WIND1	In PlantPan	Not in PlantPan
In ChIP-seq	126	2273
Not in ChIP-seq	1281	29279

Table 7 The matrix that was built for performing a Fisher exact test. The data in the table shows a comparison of the targets of transcription factor ABI5 retrieved from the Semantic Web (PlantPan interactions) against possible targets retrieved by the AraQTL webtool.

ABI5	In PlantPan	Not in PlantPan
In AraQTL	115	189
Not in AraQTL	3804	28840

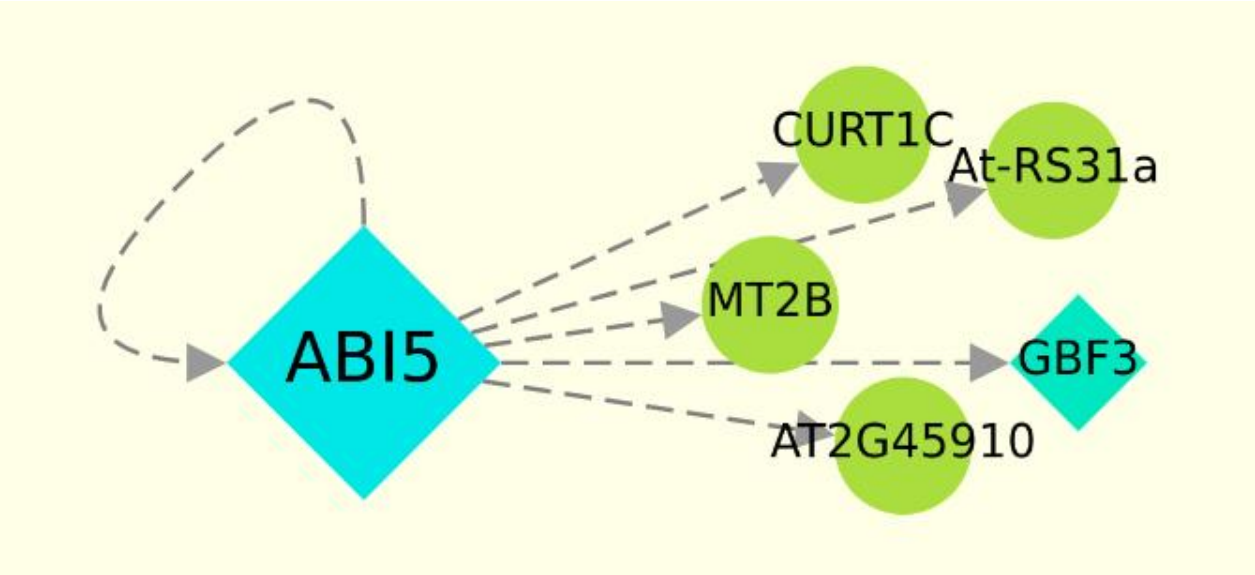

Figure 5 A TF2Network transcriptional regulatory network example for ABI5. The diamonds represent transcription factors and the circles represent the genes.

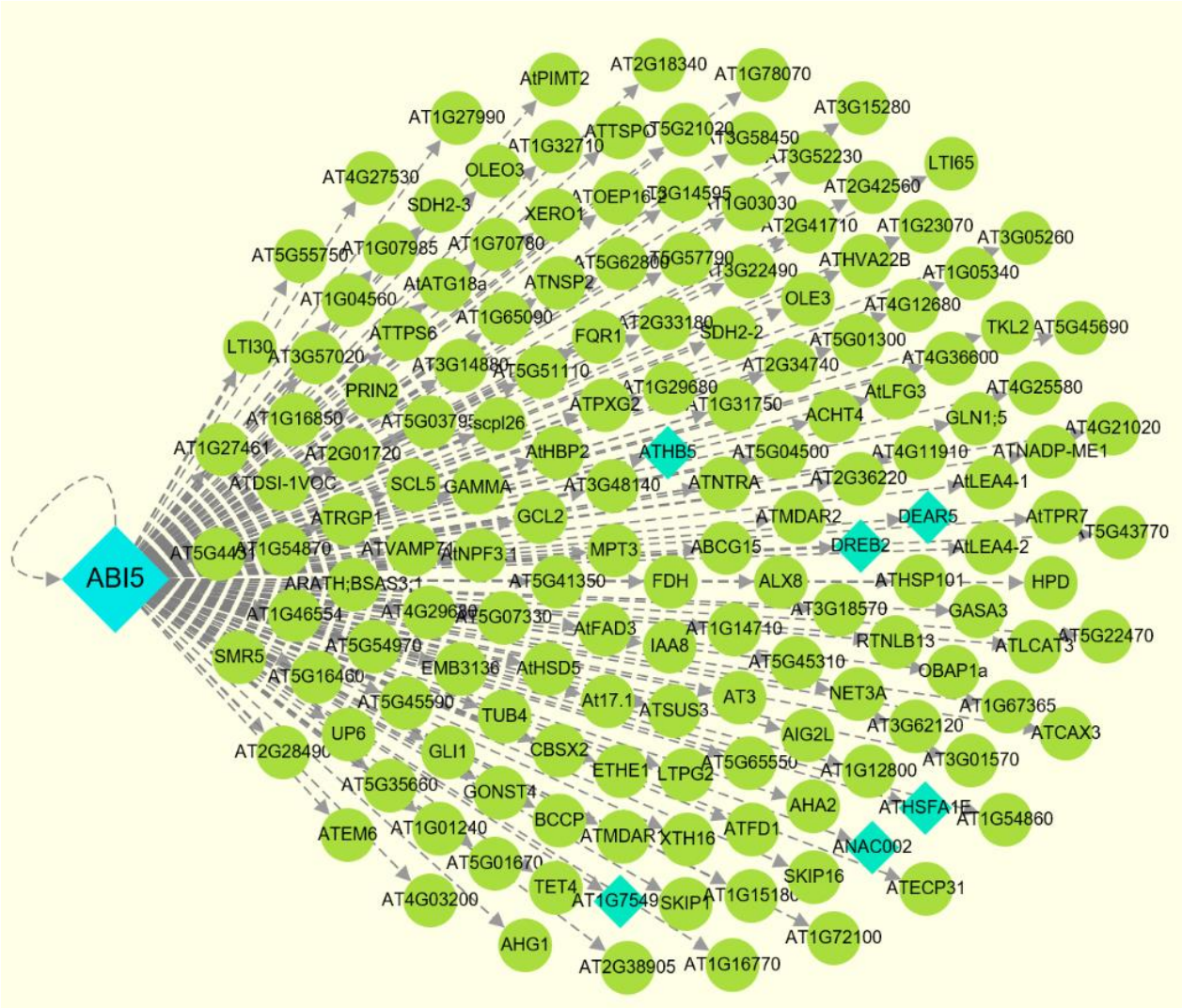

Figure 6 A gene regulation network constructed with TF2Network for the transcription factor ABI5. This network shows the was created by using the 304 potential targets for ABI5 which were predicted by AraQTL.

PWM	Rank	q-value	Hits
PWM974	1	5.42e-16	99

PWM974

source: CisBP

original id: M0254_1.01

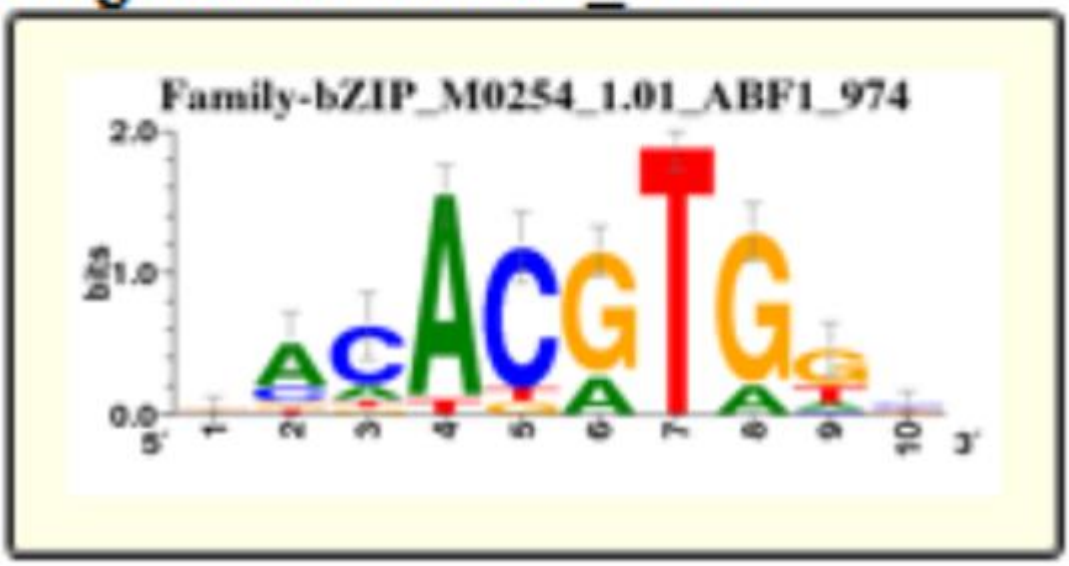

Figure 7 The motif that occurred the most when building a gene regulatory network with the 304 input genes that were retrieved by AraQTL when searching for ABI5. Here you can see the motif sequence as well as the number of times the motif was found on different targets.