BINF6112-2016S\_Final\_Project\_Report

**Genome Wide Analysis of Arabidopsis thaliana for Motif Frequency**

Satya Uppuganti, Jon Kirk

**I. Introduction**

A recent study of the *Arabidopsis thaliana* genome has identified a high frequency occurrence of a DNA sequence motif containing the core recognition site of an important family of transcription factor proteins. This discovery is of interest to the study of gene regulation, and it is possible that this sequence pattern possesses an undiscovered functional significance. We wish to first attempt to reproduce and confirm the findings of this study, and also to attempt to elucidate some of the significance of the presence of this motif by incorporating gene expression data.

**II. Background**

DNA transcription is a complex process and is highly controlled by various non-coding sequence elements. Cis-regulatory elements, such as enhancers and promoters are located upstream of genes and regulate transcription by functioning as binding sites for transcription factors. Certain patterns in the sequence are observed to be recurring in the genome and are referred to as motifs. Motifs often serve as preferred binding sites for transcription factors, and are therefore useful in studying gene regulation. Motifs are also important in other processes such as ribosome binding, mRNA transcription, and transcription termination (1.)

There is an important extensively studied family of transcription factor proteins called the DNA-Binding One Zinc Finger, or DOF, family. This family is specific to plants and has been found to be involved in the transcription of genes involved in many processes including light response and seed germination. These proteins contain a highly conserved DOF region that binds to promoters at the consensus sequence AAAG (2).

*Arabidopsis thaliana* is a small flowering plant belonging to the family Brassicaceae. This species is of very high importance in the study of plant biology and is considered a model organism, primarily due to being the first plant to have its entire genome sequenced. It has a relatively small genome of around 135Mbp, spread across five chromosomes. Because of its extensive study, genomic and transcriptional data is widely available, making it a useful subject for studying many genetic processes (3).

**III. Methods**

**I. Data download using Biopython:**

Our objective for this project is to find out the motif frequency in the A.thaliana genome. NCBI is the repository for molecular biology information. The sequences of A.thaliana can be downloaded from NCBI manually. For this project, we used Biopython(4), a python tool for computational molecular biology, for downloading the complete genome sequences using a python program. Bio.entrez module in biopython allowed us to connect to the biological databases and efetch module retrieved the records in the requested format from a list of one or more primary IDs. Biopython has sequence objects and sequence records as some of its features, that allowed us to access the sequence data from databases. All five chromosomes of A. thaliana were downloaded.

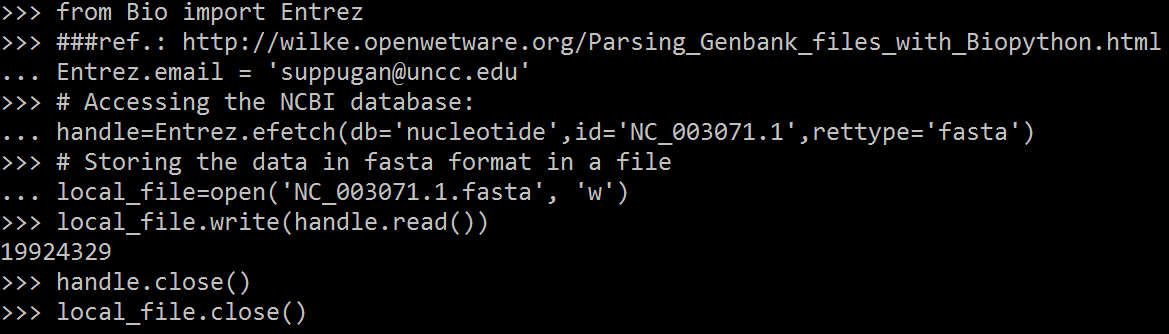


Figure 1: A Biopython program for downloading each chromosome sequences of A.thaliana from NCBI. Bio.Entrez.efetch module is used for this program.

**II: Concatenating the sequences using command line:**

Once the sequences were obtained in the fasta format, the sequences were concatenated into a single string using a Unix command “grep -v "^>" test.fasta | awk 'BEGIN { ORS=""; print ">Sequence\_name\n" } { print }' > chrms1​.fasta” that removed the line breaks in the fasta file.

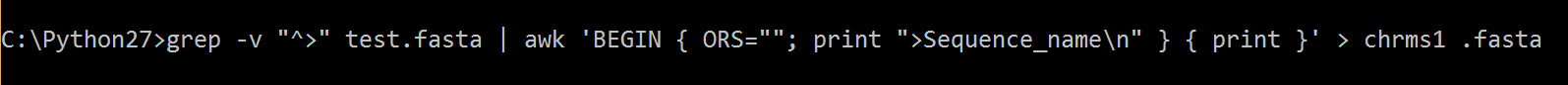


Figure 2: A unix command to concatenate the DNA sequences.

Breaking down the command, the grep function searches for the lines containing a match to a given pattern list in the input files. “-v” with a grep command returns all non matching lines i.e. it inverts the matches in grep. Awk in Unix is used for processing the rows and columns in a file. Awk has built in string functions and associative arrays. The basic syntax of AWK is awk ‘BEGIN {start\_action} {action} END {stop\_action}’ filename. The actions in the begin and the end block are performed before and after processing the file respectively and the rest of the actions are performed while processing the file.

**III. A python program for identification of motif frequency:**

A python program is used to identify the motif across the whole DNA sequence and outputs the frequency of the motifs. This code was based on the code provided in the original paper by Mehrotra et al and was adapted for our specific use(6). Our python program is unique in that it could be used for identification of any kind of motif over the sequence of any length or organism. In addition to the primary function of recording the number of motifs found, we added a functionality to the code which records the position of each motif on the chromosome. The program was run on all the chromosomes and was analysed for all four motifs templates, CTTTnCTTT, AAAGnAAAG, CTTTnAAAG, AAAGnCTTT, for varying spacer lengths (values of n) from the range 0 to 25.

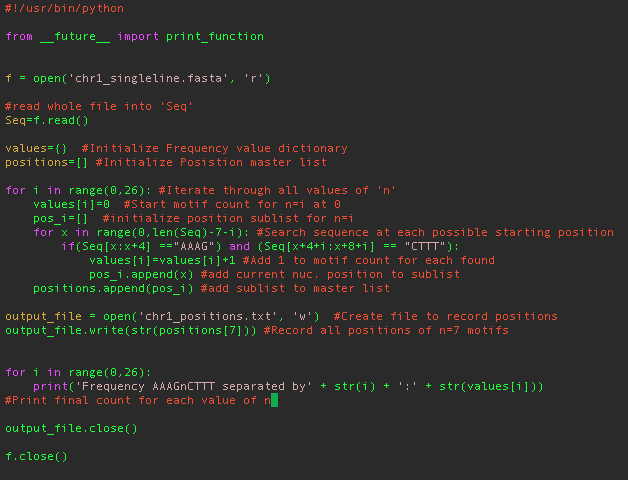
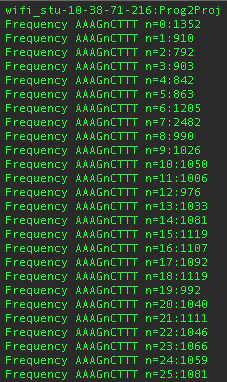
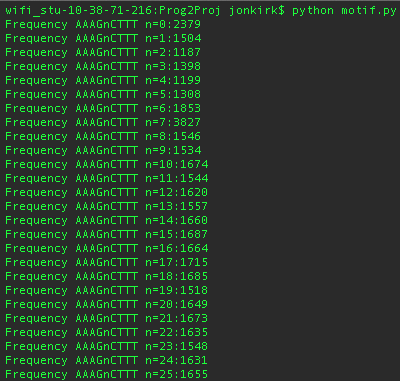


Figure 3: A python program for motif frequency detection. Every command is followed by a comment that describes its function in the program.

The motif frequency of 4 different motifs across 5 chromosomes with varying spacer length was obtained, and the data was compared to the results obtained by Mehrotra et al (6). We also wished to perform a basic statistical analysis to get a sense of the significance of high AAAG(n7)CTTT motif. We used Grubb’s outlier test, which is a statistical test for detection of outliers in a univariate data set. The test assigns a p-value which is representative of the level of extremity of the observed data point.

Figure 4: Sample output of the python program for motif frequency identification displaying the motif frequency of AAAGnCTTT motif across all spacer lengths in chromosome 1 and 2 of A.thaliana respectively.

**IV. Determine the genes with motif frequency:**

After recreating the results of the original paper, the next major step was to begin to determine which genes the motif is associated with. Because transcription factors can bind on either side of the coding region of a gene, we are investigating the upstream and downstream flanking regions of the genes. 1000 kb upstream and downstream sequences for all known genes in A. thaliana were downloaded from the TAIR database. We developed the following python program to search each 1000 kb sequence for the AAAG(n7)CTTT motif.

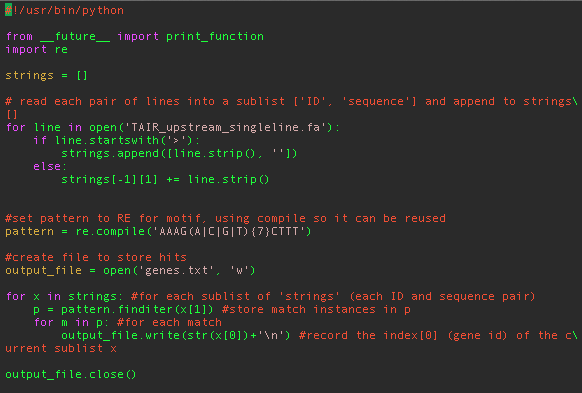


Figure 5: A python program for determination of associated genes. Each command is followed by a comment that describes its function in the program.

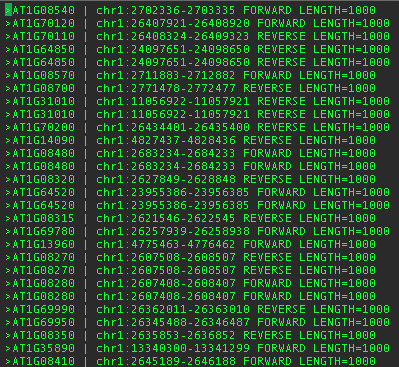


Figure 6: A portion of the output of the python program displaying the gene ID and the gene location.

**V: Results and Discussion:**

After using the python program to record the number of all motifs on all chromosomes, we found that all of our numbers matched up perfectly with the results obtained in the original paper.

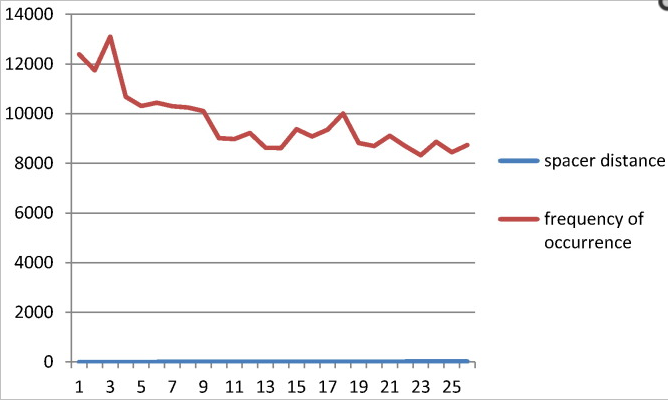
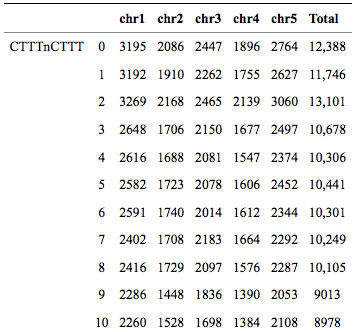


Figure 7: Sample of results obtained by Mehrotra et al (6)

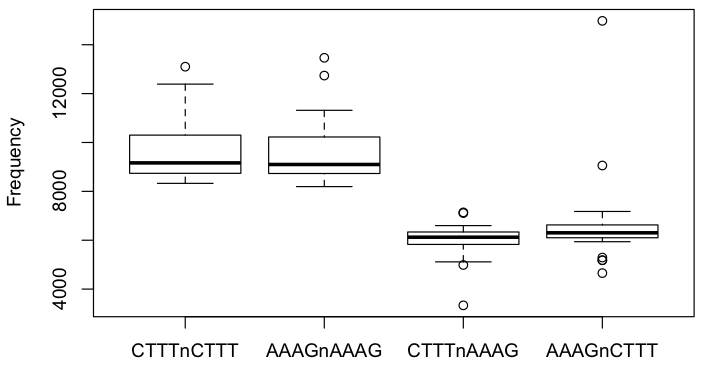


Figure 8: Plot of our compiled results showing the motif frequencies of 4 different motifs using Grubb’s outlier test in R program

**V.a.Statistical Analysis:**

In R, we used the Grubbs’ outlier test to assess four data sets - the genome totals for frequency of values 0-25 for each of the four motif templates - and the test assigned a p-value to the highest point in each set (the motif of value n seen with the highest frequency.) The frequency of the motif AAAG(n7)CTTT, our motif of interest, received a p-value of 2.98e-09.

**V.b.Motif frequency of A.thalina vs. O.sativa:**

Our objective for comparing the two plant species’ genomes for motif frequency was to find out whether the occurrence of DOF motif AAAG(n)CTTT is specific only to A.thaliana or it is widely spread across the plant kingdom by evolutionary processes. In the case that the motif was found in O. sativa, we were also interested in the prospect of being able to compare the function and gene relationship of the motif between the two species. However, the results indicated no specific occurrence of high frequency of motifs across the seventh chromosome in O.sativa.

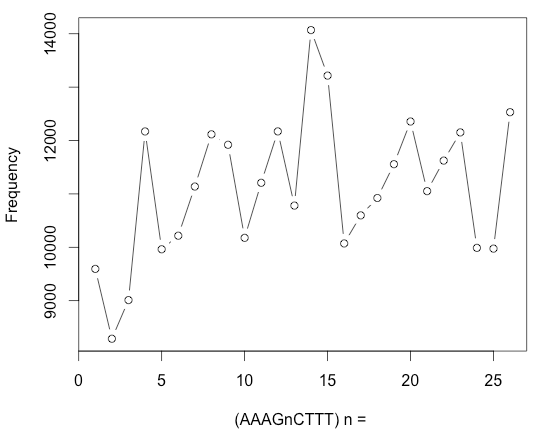


Figure 9: A plot displaying total AAAGnCTTT motif frequency in O.sativa genome.

We were able to successfully confirm the results obtained by Mehrotra et al., and used a statistical analysis to verify the abnormally high frequency of AAAGN7CTTT motif in A.thaliana. In our assessment of the frequency of this particular motif in O. sativa, we found that there is no preference shown. In order to obtain a list of putative genes associated with this motif of interest, we developed a python program to search for it in the upstream and downstream flanking regions of all known genes in A. thaliana.

**VI. Functional Analysis:**

In order to begin assessing possible functions of the motif preference, we looked at whether there was a general preference for the N7 motif in gene flanking regions relative to its own background frequency, in comparison to the other values of N. The Python program which searches the flanking regions for motifs was changed so that it searched for motifs containing each value of N consecutively and created output files of the genes identified for each.

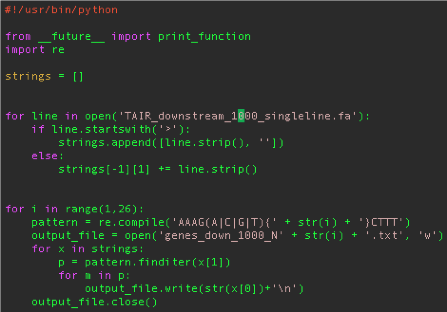


Figure 10: A python program for searching the flanking regions for motif frequency

For each motif with a particular value of N, the total number of motifs found in either gene flanking region was taken as a percentage of the total number of the same motif found anywhere in the genome. This analysis determined that all of the motif varieties, including N7, were encoded into the gene flanking regions at a rate of about 50-60%. So while the total number of N7 motifs found in gene flanking regions is higher, there does not appear to be a general preference for it in these regions because it is encoded at about the same rate, and the higher frequency can be explained by its higher background frequency.

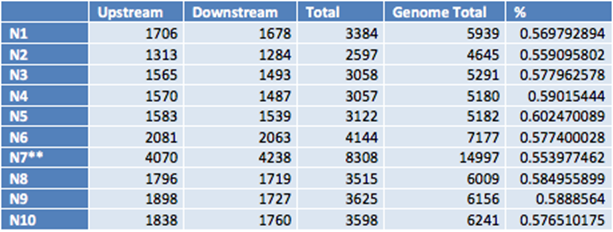


Figure 11: Calculation of motif background frequency

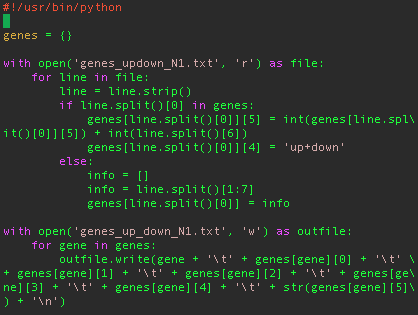


Figure 12: A python program for creating the rows of Gene ids that are present in both up and downstream regions with motif frequency

**VI.a. Database for functional analysis:**

A comprehensive database using SQLite manager (version 3.8.7.4) of Mozilla Firefox version 19.0 was used for managing and analyzing the genomic tables with huge number of rows. SQL statements were to work with the queries for functional analysis of the genes of A.thaliana. The database was imported with all the gene files that were downloaded from the NCBI and also with the gene ontology (GO) and plant ontology (PO) tables that were downloaded from the TAIR website. The dataset that was obtained from our python program that gives information about the motif frequency was also imported.

Gene ontology is a bioinformatics initiative that is used for unifying the genes with gene product attributes for all the plant species. The GO defines classes used to describe gene function, and relationships between these classes. The functions are classified into three concepts: molecular function - that describes the molecular activities of gene products, cellular component - that describes where the gene products are active and finally the biological process - which gives information about the pathways and larger processes made up of the activities of multiple gene products. Using these data, the TAIR gene IDs were mapped with NCBI IDs to obtain more usable information about the genes such as gene descriptions, symbols, synonyms etc (Fig 14).

PO is a collection of ontologies developed by the Plant Ontology Consortium. These were used for describing the anatomical structures and growth and developmental stages across Viridiplantae. The genes were successfully associated with the TAIR ids using our database (Fig 13). Unfortunately, no conclusion about the genes being associated with a particular plant ontology could be made as each and every gene has multiple PO terms associated with it which made it difficult to group the genes based on the PO terms. However, this information can be used for future purposes if we want to continue our research with PO.

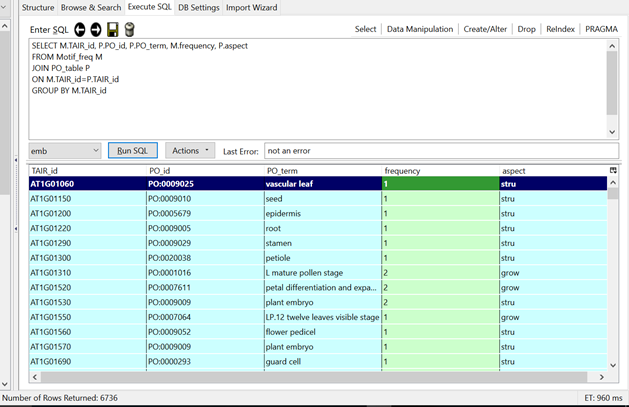


Figure 13: A screenshot of the database displaying the genes associated with the PO terms among the motif frequency dataset.

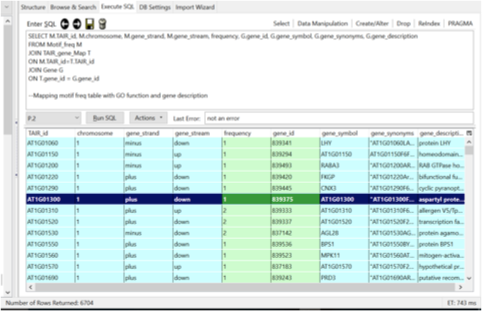


Figure 14: A screenshot of the database displaying the gene descriptions (symbol, synonyms, description) associated with the motif frequency dataset.

**VI.b. Gene Enrichment Analysis:**

Our goal for using ontologies is to conduct the gene enrichment analysis to assign biological meaning to the group of genes and to assess whether some subset of the group shows significant overrepresentation of some biological characteristic[. The ‘group by’ function in the SQL statement was used to obtain the total number of all the genes that were associated with a particular GO function in both the whole genome and the dataset tables (Fig 15,16). These numbers were used in the following formula to obtain the enrichment factor (Fig 17).

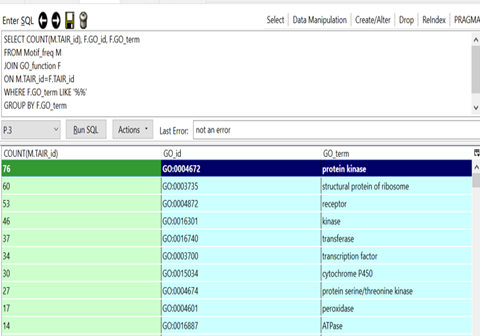


Figure 15: A screenshot of the database displaying the total count of the genes associated with each of the GO functions among the genes in the motif frequency dataset.

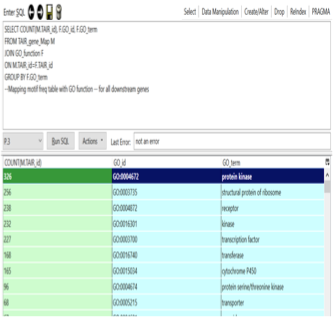
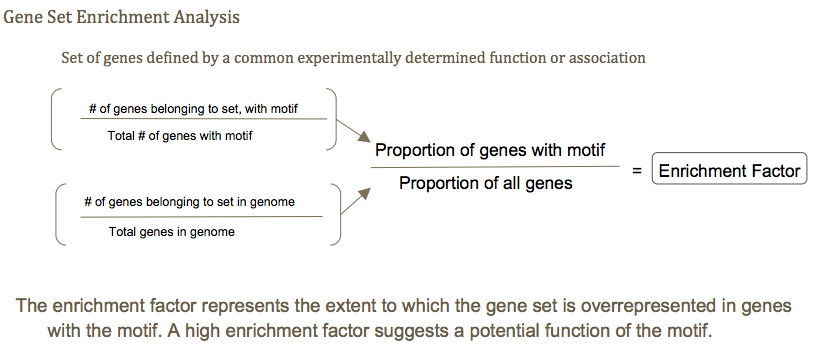


Figure 16: A screenshot of the database displaying the total count of the genes associated with each of the GO functions among the whole genome of A.thaliana.

Figure 17: Formula for obtaining the enrichment factor

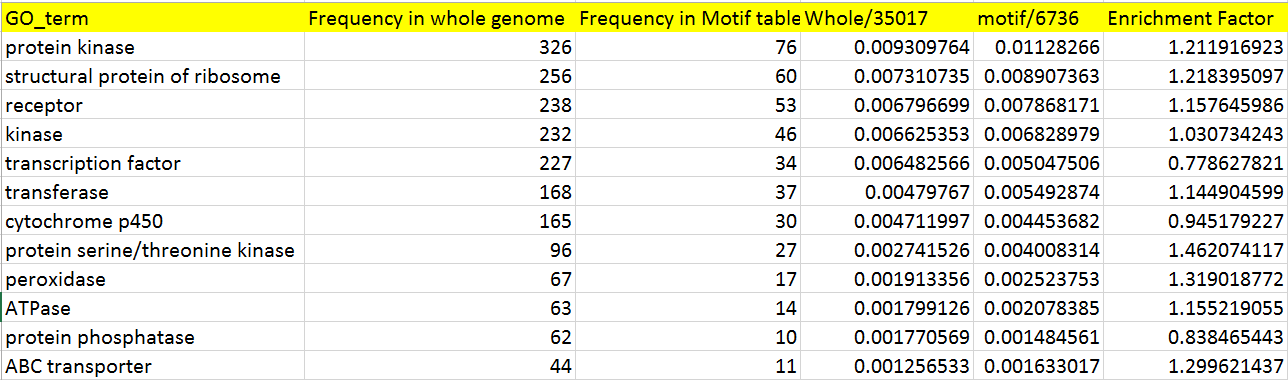


Figure 18: Calculation of enrichment factor

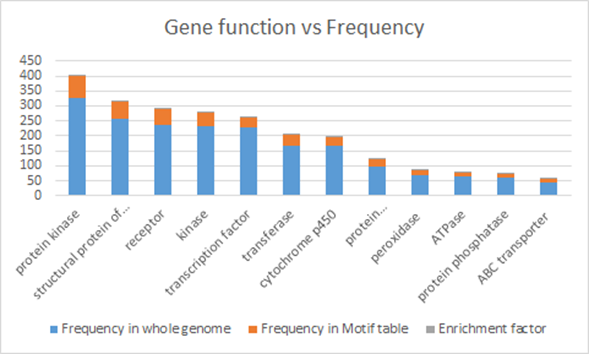


Figure 19: A plot displaying the frequencies of gene functions among the whole genome and the motif frequency dataset for enrichment analysis.

**Functional Analysis Results:**

From the enrichment factor results, it was found that though most of the the enrichment factors were a whole number, these values were not significant enough to conclude that a subset of genes were associated with a specific function.

**VII. Conclusion**

Our main focus for this project was to find if the AAAGN7CTTT motif was highly repeated across the whole genome of A.thaliana using a python program and then use the same program to compare our results with that of O.sativa. We successfully confirmed that AAAGN7CTTT exists in significantly higher frequency in A.thaliana but no such significant pattern was observed in case of O. sativa, though the DOF protein family which has this motif is evolutionarily conserved between these two species. We also discovered the genes associated with motifs of interest by locating in upstream and downstream region and determined that the N7 motif is not preferentially encoded in gene flanking regions. By building a comprehensive database on A. thaliana gene data, we performed functional analysis by calculating the enrichment factor for genes with motif frequency and looked to associate genes with possible functions. But with the very low enrichment factor values obtained from the results, no association could be made. However, using the GO tables in our database, we were able to identify the set of genes that had a particular GO function term allowing us to understand various roles of the genes in the biological processes. Finally, we assessed PO associations of the genes with roles in developmental, structural and growth of a plant.

**VIII. References**

1. D’haeseleer, P. What are DNA sequence motifs? Nature Biotechnology 24, 423 - 425 (2006).

2. Noguero M. et al. The role of the DNA-Binding One Zinc Finger (DOF) transcription factor family in plants. Plant Sci. 209, 32-45 (2013).

3. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana.* Nature 408, 796-815 (2000).

4. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, Peter M. Rice: Nucleic Acids Research **38** (6): 1767–1771 (2010). [doi:10.1093/nar/gkp1137](http://dx.doi.org/10.1093/nar/gkp1137)

5. <https://en.wikipedia.org/wiki/Grubbs%27_test_for_outliers>

6. Genome wide analysis of Arabidopsis thaliana reveals high frequency of AAAG N7 CTTT [motif](http://www.sciencedirect.com/science/article/pii/S2214540014000346) [R Mehrotra](https://scholar.google.com/citations?user=vJzFrpsAAAAJ&hl=en&oi=sra), V Jain, [C Shekhar](https://scholar.google.com/citations?user=11kNL3wAAAAJ&hl=en&oi=sra), S Mehrotra - Meta gene, 2014 Dec; 2: 606-615.

7. <http://www.cs.tau.ac.il/~rshamir/ge/09/scribe/lec14a.pdf>

**Supplementary files:** The following are the python program files used for our project.

* Motif.py
* getgenes.py
* up\_and\_down.py