Motif Assessment of Vernalizing Grasses

Isaac Racine
Department of Biology
and Department of Data Science
University of Vermont
December 6, 2020
isaac.racine@uvm.edu

ABSTRACT

Grasses contain several cereal crops which provide the largest amount of food security around the world. Many species of grasses must undergo a process of vernalization, enough cold temperature exposure and shorter amounts of daylight to promote proper flowering in the spring. With the climate warming it is vital to understand what genetic elements are regulating this process in hopes of being able to make mutant crops that no longer need to undergo vernalization or so that they can be well understood for indoor and vertical farming. Many motifs, conserved sequences across a clade of closely related species, are often places that transcription factors or other regulatory elements can bind to regulate a given gene. A motif analysis was preformed on representative species of the grass family to identify cis-region, regions within the gene but not protein coding, motifs that could be responsible for the regulation of the two important genes involved in vernalization, FUL1 and FUL2. After identifying these cis-region motifs a model was construncted to see which were most valuable in identifying core *Pooideae* species and relatives, a small clade within grasses containing wheat and barley. The motifs used for building the model are seemingly important in the evolution of the core *Pooideae* and relatives. However, more analysis should be performed to help identify novel motifs or motifs related to regulating proteins. Either of these types of motifs may hold the answer to how vernalization is being regulated, which will be beneficial to future experiments wishing to make genetic mutants..

Keywords

vernalization, motif identification, cis-regulating regions, *Pooideae*, *Poaceae*, cereal crops

1. INTRODUCTION

Cereal crops are the major source of nutrients to provide food security to many communities and cultures around the world. Many of these species belong to family *Poaceae*, also known as grasses [23]. Since these cereal crops provide the most amount of food security to people around the planet it is important that these crops continue to grow with high yields. As the climate warms this poses a threat on the yield and security of these crops. This is due to many of these species needing to undergo a process called vernalization to flower properly in the spring. Vernalization, also known as overwintering, promotes flowering after a plant has endured a sufficient time of cold temperatures [6]. Another contributing factor of flowering time is photoperiod. A decrease in the photoperiod along with extended cold exposure can induce prime flowering responses in plants that must undergo this treatment [6]. Proper flowering time results in higher yield for these cereal crops. Thus to combat the warming climate the genetic pathway controlling these species' response to vernalization is hoped to be better understood to make sure cereal crops can remain secure to the food structure of humanity. To better understand this a model was constructed using a clade within *Poaceae* so that the species are more related, core *Pooideae* and relatives. This allows for specific genetic regions to be more easily identifiable as important in regulating vernalization for later evolved cereal crops. What motifs are most characteristic of the core *Pooideae* and relatives?

The vernalization pathway has been well studied. The story starts millions of years ago when a duplication event of a vernalization gene resulted in paralogs, FUL1 and FUL2. Both of which have influential roles in several grasses responses to vernalization, even those grasses that do not require vernalization [3]. However, all species examined are ones that must undergo overwintering to flower correctly. The FUL1 (VRN1) gene is thought to cause the induction of flowering genes, along with genes measuring the photoperiod, amount of sunlight [23]. FUL1 is naturally activated, so it usually is repressed by another gene in the pathway, FUL2 (VRN2), during times of non-reproductive growth. The FUL2 gene in turn is repressed by environmental signals such as short days and extended periods of cold temperatures, thus allowing for the promotion of flowering during the springtime [1]. FUL3 (VRN3) is an activator of the FUL1 gene and is also usually repressed by FUL2. FUL3 is induced by longer days, as winter turns to spring, and by the termination of repression by FUL2. This shift from vegitative to reproductive phase in overwintering plants is irreversible, so the response to these environmental signals are very sensitive [21]. Since FUL3 is not as influential in the pathway only FUL1 and FUL2 will be reviewed.

Although this response pathway is highly regulated little is known about what is actually regulating the expression of these genes. Of course the environmental signals are the main signals, but which proteins these environmental signals interact with to cause the change in response is lacking. However, some prior research highlights that many *Brachypodium*, a genus in the *Poaceae* family and the close relative to core *Pooideae*, have many regulatory elements, regions that regulate a genes production, in cis-egions of the gene reviewed [15]. These sites usually act as transcription factor binding sites. Thus, when thinking of how to narrow the search it seemed natural to focus on the cis-regulatinig regions of the vernalization genes. Although the search is for all of *Poaceae* the group *Pooideae* and close relatives (*Brachypodium*) within the grasses were further investigated. This is because this subset contains many cereal crops, while being more closely related evolutionarily. This smaller scope allowed for more precise results of which motifs are important in regulating the vernalization pathway, even for a small subset like *Pooideae*, since this is an explorative experiment. The representative species for core *Pooideae* and close relatives are H. vulgare, *B. stacei* and *B. distachyon*.

To explore the possibilities 11 representative species of the ingroup, *Poaceae*, had the full gene sequences of both FUL1 and FUL2 analysed. Additionally, two outgroup species whose ancestors evolved before the FUL duplication event are analyzed but only have one gene, FULLIKE. This analysis focused on motif discovery using the MEME algorithm [2]. Motifs are stretches of DNA that are conserved across several species. Due to the nature of motifs the exons, protein-coding regions, were ignored because of their high conservation across related species and unlikeliness of having regulating capabilities. The organizing of what motifs were possibly influential was done using a pipeline that determined the region each motif was found in for a given gene and species. The only motifs reviewed for importance were those found in cis-regulating regions, as these regions are thought to have the greatest impact on regulation, level of protein sunthsis, of the vernalization pathway. Although all introns are considered cis-regions only the first intron was reviewed. This first intron within the FUL1 and 2 genes is thousands of bases long and huge comparatively to the other introns. This long intron, the promoter and 5' untranslated region likely have motifs hidden within that hold the answer to the regulation of these genes. Motifs were then used to construct a logistic regression model to identify the likelihood of being in a core *Pooideae* or relative. This model was made with motifs picked by a feature selection method to prevent overfitting of the model. The five most important motifs in predicting this were found and can be used in later experiments to create genetic mutants.

If the motifs important in regulating vernalization are found then it will allow for humans to respond to climate change while maintaining food security. This is because even if the climate continues to warm these mutants can be modified to undergo no or less wintering. The yield in crop biomass is reduced and the flowering time is delayed when not enough wintering has occurred or if it is sporadic [6, 7]. Although climate change itself is important to combat, there will likely be nobody to combat this with food scarcity sweeping across the globe. In an attempt to provide sustainable food the experiment performed tries to find insight into the underlooked regions of vernalization genes, intergenic cis-regulating regions.

2. RELATED WORK

Unfortunately little research has been done in the context of motif identification within the vernalization pathway. The main motif known to exist in all the vernalization genes across species is the CAr-G motif, as this motif is important in development for plants. However, there are varying stances on the importance of this motif in the regulation of vernalization [16, 24]. The few studies that have reviewed the role of the CAr-G motif have been limited to wheat species, rather than an entire clade within the grass family. Other studies have found an influence in vernalization response when the 5' untranslated region was mutated [8]. Additionally, the first intron of the vernalization genes is very large, several thousand base pairs long. Some studies have shown that deletions within this region can greatly influence the response to vernalization, at least in barley and wheat [9]. So there is a great possibility that motifs within the first inton may be greatly contributing to the regulation of the vernalization genes.

In general there is likely a colescent influence from several motifs within these cis-regulating regions that prior studies have already explored and determined is contributing to the regulation of vernalization. Many genes are not simply regulated with just one binding site, transcription factor or regulating sequence. Instead many genes are regulated by several binding sites, regulating proteins or other genes, making the phenotypic expression dependent on several variables. These genes are known as complex genes. Thus, trying to understand which regions, motifs, within the gene are contributing the most to the regulation of the gene is what is important. Luckily, a prior study had a nice pipeline for this examination [17]. The researchers examined cis-regulating motifs in another gene, but the methodology can be applied to the vernalization genes. Although not all of their pipeline was used, the backbone for the pipeline used in this exploratory experiment was based on theirs [17].

3. METHODS

3.1 Obtaining Gene Sequences

The following processes were done using Python version 3.8 [18]. The data, gene sequences, for these genes were obtained from Phytozome, an online database for several species of plants' genomes(Phytozome:https://phytozome.jgi.doe.gov/pz/portal.html). The Phytozome website offers a useful query builder for web scraping called Phytomine. From Phytomine the gene sequence for each gene was scrapped, including 2000 bases upstream of the 5' untranslated region for each gene. These 2000 bases are representative of the promoter region for each gene. Along with the gene sequence each splice variant transcript for each species's gene sequence was scraped along with each splice variants corresponding protein coding sequence. The number of splice variants a gene has is random, however one transcript is usually made more often than others, the primary transcript. From the many sequences two different file types were constructed. The first type of files contained only the genomic sequences and were saved to files with just FUL1, just FUL2 or both FUL gene sequences. Additionally, each of these three types of files had either no outgroup sequence or both outgroups sequences for their FULLIKE genes present. This created a total of six files that are later ran in MEME. Additionally, the other types of files created were to be used for alignment. Each of the genes' splice transcripts were saved to their own fasta file with their respective gene and protein coding sequences. Since splice transcripts come from the same gene sequence those gene sequences should be the same sequence and length for each of its respective splice transcripts. These files were then each individual aligned, as discussed later. Lastly a file containing all of the protein coding sequences was scrapped to be used for phylogeny building.

3.2 MEME Tests

To perform the following processes the BioPython package was downloaded and used in Python [5, 18]. Each of the six files containing different combinations of just gene sequences were passed through MEME. MEME has several useful parameters to set for finding motifs. For each test only the top 15 motifs were recorded and the maximum sequence length was 12 bases. 12 bases was picked as a length because it seemed like a representative length for transcription factors to bind or for regulating transcription [22]. Additionally, each file had two tests performed with the above parameters. The two types of tests are ZOOPS and ANR. ZOOPS stands for zero or once per sequence, while ANR stands for any number of repeats. The differences between these tests is that ZOOPS only returns motifs that occur at most once per sequence within a group of sequences. The ANR test will return motifs that occur any number of times per sequence, including motifs that occur zero or once. Both of these tests were run to better encapsulate the possible motifs regulating vernalization response. So in total 12 MEME tests were performed, two tests for each of the six files, and all 12 test results were saved in XML format.

From the 12 test files a large dataframe was constructed containing each instance of a motif that was identified as one of the 15 most conserved. This means that the gene identifier, motif length, motif start position in its respective sequence, motif sequence and motif's consensus sequence could be extracted from the files using BioPython functions [5]. Then feature engineering was performed so that each gene and species a motif was in was represented in its own column, along with the specific MEME test each motif originated from.

3.3 Region definer and motif region identification though feature engineering

The several files containing each splice variant and its corresponding genomic and protein coding sequences were all aligned using MAFFT [13,14]. Since all of the three sequences in a given file are from the same code, gene sequence, there should not be any gaps in the gene sequence. Gaps can occur in sequences when aligned because of their varying length or variability in base position. The other sequences, splice transcript and coding sequence, should have gaps where the promoter is and introns as these regions are not present in those sequences. This allowed for a function to be written that determined the start and end position of each region for a given splice variant. All of these start and end positions of each region within each splice variant of each gene were saved to a large dataframe. Since the start and end positions were found this allowed for the motifs to be placed into regions because each motif's start position is known. A new feature that said which region a given motif occurrence was found in for that given gene was added to the motif dataframe.

The function that found what region a given motif was in did so by matching the motif to the proper splice variants by the gene identification. To catch for any errors if a start position for a motif was not in range of any of the regions for a given splice variant a NA was recorded instead. The number of motifs for a given region was displayed in a barchart to explore the results of the function written (Figure 2). The NA column was rather high so more exploratory analysis was done. Barcharts grouping NAs by species and splice variants were also constructed. The only species containing NAs were species that had more than four splice variants. After reviewing the raw data that was scraped from Phytomine it was determined that Phytozome actually returned different gene sequences for some of the splice variants. This should not be the case considering each splice transcript for a species can only be processed from the same gene, and thus can not have varying lengths in the gene sequence.

3.4 Parsing out faulty data

To fix the problems of motifs with a region of NA only the primary splice transcripts were used. The primary transcript had no accuracy problems when scraping from Phytomine and is the most important in terms of protein synthesis. Thus all of the transcripts that were not the primary were removed from the scrape. After rerunning the functions to define the start and end position for each region and find the corresponding region a motif is in there were not any motifs placed in NA regions. This data was then used for the rest of the experiment.

3.5 MrBayes

The protein coding sequence of each of the primary transcripts were aligned using MAFFT [13, 14]. This alignment was then passed to MrBayes [11, 19, 20]. MrBayes is an application that uses Bayseian inference to construct a phylogenetic tree. The application also uses a bootstrapping approach to increase the statistical power of the results. The phylogenetic tree obtained was in newick format. Any phylogenetic tree constructed was done so using functions from the ete3 package in Python [12, 18]. Since this phylogeny is already known the results were no surprise. However, a function was written to print the genetic diagram for each gene next to their corresponding gene identifier. To do this a data frame of just the motifs in the cis-regulating regions was made. The function read through the cis motif data and the region data to produce the visual. The function colored in the three most important cis-regulating regions along the gene in this context and added bars along these regions that were colored differently. The differently colored bars were representative of the different motifs identified. To avoid overcrowding of motifs on a single tree two trees were made, one of motifs from the ZOOPS tests and another with motifs from the ANR tests.

3.6 Modeling

A logistic regression model was constructed to predict which motifs were most important in identifying core *Pooideae* species and relatives. Originally a PCA model was constructed, but since the actual motif sequences needed to be known for future experiments this model was abandoned. To construct the logistic regression model the motifs characteristic of core *Pooideae* and relatives had to be determined. This was done by feature selection, where each unique motif from all of the 12 MEME tests were treated as possible features. To determine which motifs, features, to select for some more features first had to be engineered. For each motif that was in a core *Pooideae* and close relative species a 1 was added, otherwise a 0 was added to a new column in the cis motif dataframe called "Pooids". To determine if a motif was present in all core *Pooideae* and relatives an iteration through the "Pooids" column was performed. If a unique motif was in all core *Pooideae* and close relatives then that motif for each of those entries received a 1, and all other motifs received a 0 in a new column called "All Pooids". So even if a motif was in two of the three species that motif still received a 0. This column, "All Pooids", was then used as the dependent variable in the feature selection and logistic regression model.

Remember that only motifs in the cis-regulating regions were used for feature selection and modeling. The column containing the consensus sequence of each motif was then one-hot encoded so that they could be used as features. A test and train split of 70% and 30% was performed. From this a feature selection using a chi-square approach was performed on all of the unique motifs within the training data. This was due to the new motifs features being binary and the fact that some motifs would not be as good at predicting core *Pooideae* and close relatives than other motifs. So in order to prevent motifs that may be within all species or just totally independent of being in all core *Pooideae* and relatives a chi-square approach was decided on to weed out the most dependent and thus characteristic motifs of these species. The feature selection only considered the one-hot encoded motif variables and none of the other variables. The five motifs with highest chi-values from feature selection were chosen from the selection model. Five motifs, features, were selected because this seemed like a proportional amount of motif identifiers compared to the number of all unique motifs identified, over 120. These five motif features were then also used to construct the logistic model where the "All Pooids" column was used as the dependent variable. The test data was then used on the training data's model to determine the accuracy of the model. Depending on the accuracy of the model could have allowed for more or fewer features to be selected for. Each feature was individually modeled in predicting in being in "All Pooids" and then stacked on eachother in a regression plot. This allowed for examination of the trends present in the model.

3.7 Final Tree Construction

Again the ete3 package was used in phylogeny construction [12]. A final phylogenetic tree was constructed, but the only motifs presented in the visual genetic diagram were the five "most important" in characterizing the core *Pooideae* species and close relatives. These motifs were the five selected from the feature selection process above. The tree showed the five motifs for all genes in the tree so not all of the motifs were foun in the three primary cis-regulating regions purily because of the variation in code and location that evolution produces.

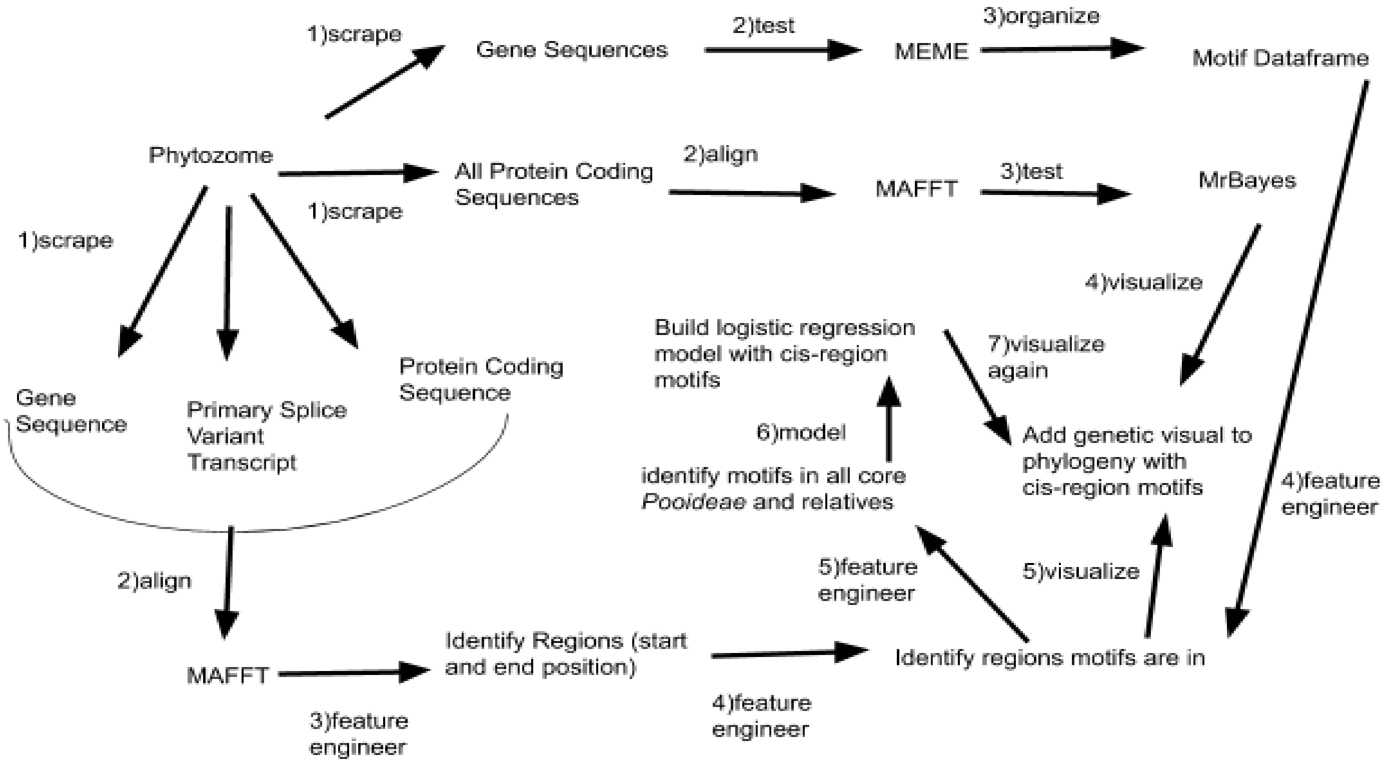

1)scrape
Gene Sequences
2)test
MEME
3)organize
Motif Dataframe
Phytozome
All Protein Coding Sequences
2)align
MAFFT
3)test
MrBayes
1)scrape
1)scrape
4)visualize
Build logistic regression model with cis-region motifs
7)visualize again
Gene Sequence
Primary Splice Variant Transcript
Protein Coding Sequence
6)model
Add genetic visual to phylogeny with cis-region motifs
identify motifs in all core *Pooideae* and relatives
4)feature engineer
5)feature engineer
5)visualize
2)align
Identify Regions (start and end position)
MAFFT
Identify regions motifs are in
3)feature engineer
4)feature engineer

Figure 1: Shows the pipeline used to obtain the results gotten. The arrows show the flow of how processes must be completed. Most of the data was stored in data frames along the pipeline to allow for organization and faster analyses.

4. RESULTS

From the exploratory experiment performed there were many motifs identified as possible candidates for regulating transcription in cis-regulating regions. From the pipeline (Figure 1) a model was able to be constructed for predicting the most important cis-region motifs for identifying core *Pooideae* and close relative species. However, getting to the results of the most important motif required a lot of data cleaning and feature engineering along the way. Luckily some phylogenies Lalong with genetic diagrams were able to be constructed from the data and be improved by the model.

4.1 Exploratory Data Analysis and Phylogeny Building

Scraping all the splice variants from Phytozome and determining the region each motif was in for each splice variant given its gene identity showed problems. The number of motifs found in the NA region was very high (Figure 2.1). This caused more exploration to be done. A bar chart of motifs in the NA region were grouped by species and splice variant (Figure 2.2) As a result the splice variants that were not the primary splice variant were dropped which eliminated all NAs. Using the motif start position and the known gene length allowed for the visual to be added to the phylogeny (Figure 3). This emphasized just how many motifs were in the first intron and cis-regulating regions(Figure 2 and Figure 3).

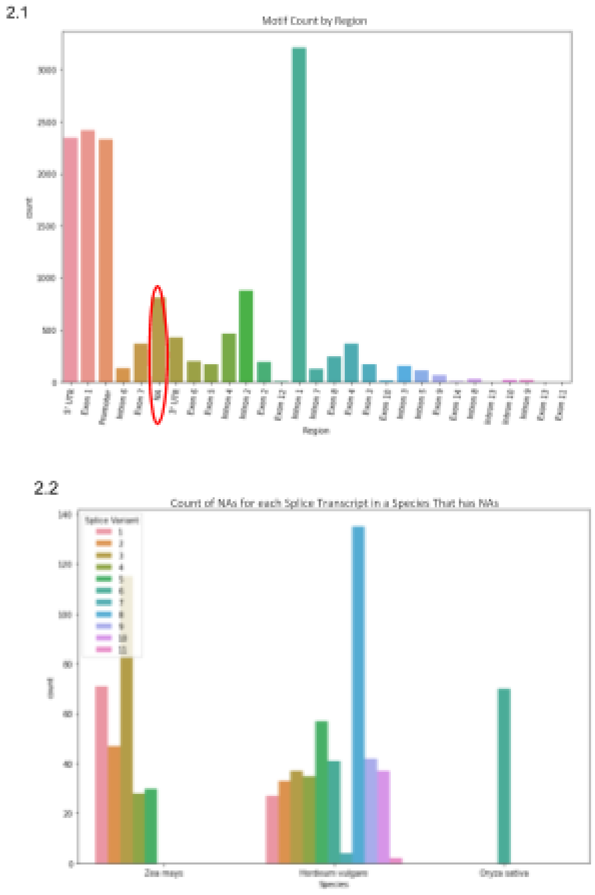


Figure 2: 2.1) Shows the counts of motif by region when all splice forms were considered. This graph was constructed during the exploratory data analysis to see if the motifs were being placed in correct regions. Specifically, the NA column should be minimal if done correctly. The NA column is circled in red. 2.2) Shows that species with motifs in a NA region all had NAs in several splice transcripts and that all of these species had several splice variant transcripts.

Together all of the features are used in constructing the model for predicting which motifs are important for characterizing these species.

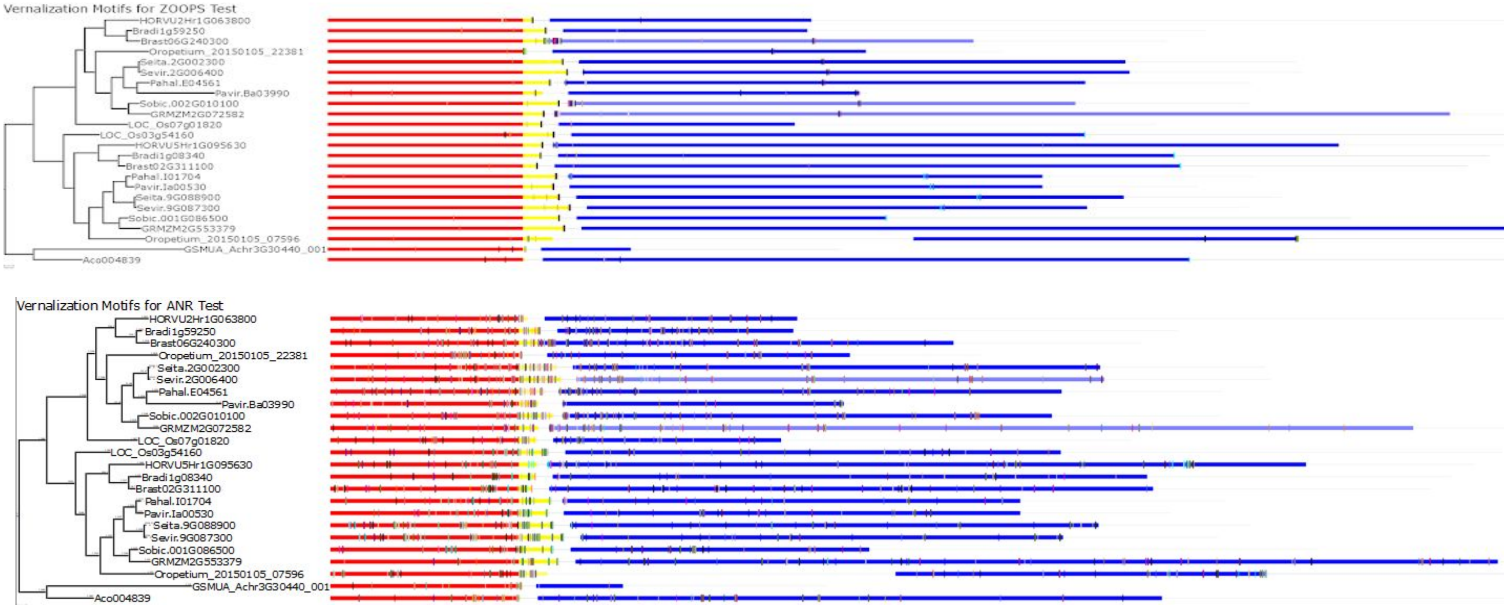


Figure 3: Shows the phylogenetic relationship of all genes sampled. The phylogeny was constructed using MrBayes. The genetic visual to the right shows where the cis-regulating motifs are. The red region represents the promoter, the yellow represents the 5' untranslated region and the blue region represents the first intron. Each unique motif has its own color and is represented by vertical bars along a gene visual. The first tree shows the motifs from the ZOOPS tests and the second tree shows motifs from the ANR tests. The ANR tree has more motif occurrences, colored bars, because the test does not limit motifs to one occurrence.

4.2 Feature Importance and Modeling

The logistic regression model using the five selected motifs to predict if they are useful in characterizing core *Pooideae* and relatives or not was constructed. Since the chi-square feature selection was implemented on the motifs the five highest chi values were presented with their corresponding standard error bars, p-values and the motifs corresponding consensus sequence (Figure 4). All of the p-values for each feature, motif, are significant. To explore the spread of the important motifs across genes a bar chart was constructed (Figure 5). Most of the motifs are occurring in FUL1 genes and furthermore coming from FUL1 MEME tests. The logistic regression for each individual feature, motif, was plotted one on top of another to see each motif's individual contribution to the model (Figure 6). Thus, the trend lines are for predicting if a motif is in all core *Pooideae* and relatives or not in them. Two main trends emerged, with some motifs being important at characterizing not core *Pooideae* and close relatives. The other trend consists of motifs that are important in identifying core *Pooideae* and close relatives.

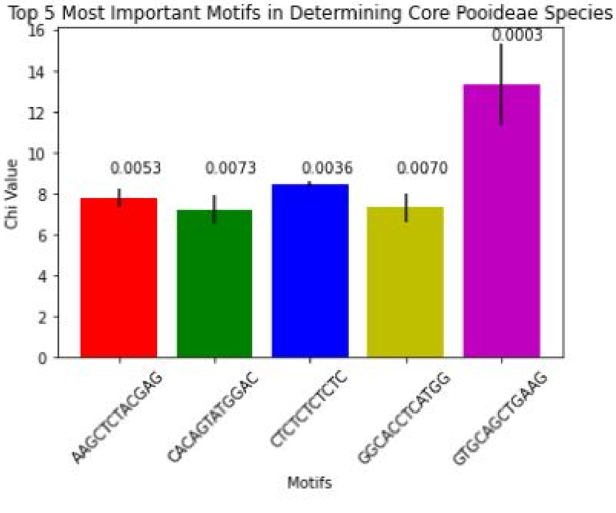


Figure 4: Shows the motifs with the highest chi values after a chi-squared feature selection approach of all of the unique motifs. The standard error bars, 95% confidence intervals, are also presented. Above each error bar shows the p-value of each feature, motif, from the feature selection test. Along the x-axis shows the consensus sequence of each of the most important motifs in identifying core *Pooideae* and relatives.

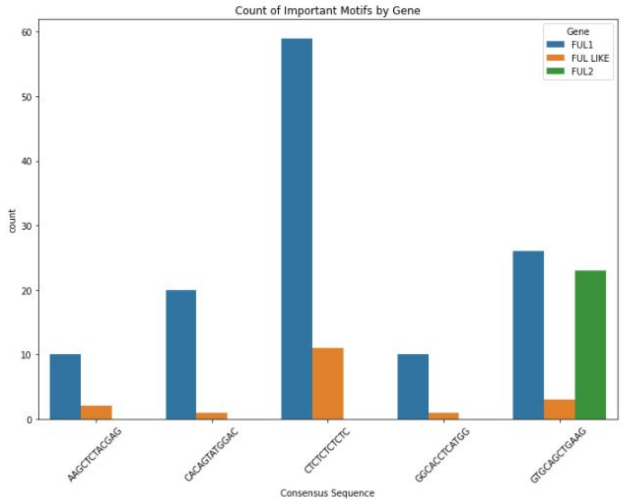


Figure 5: Shows which gene each of the top five feature selected motifs are present in. FUL LIKE genes are from the two outgroup species.

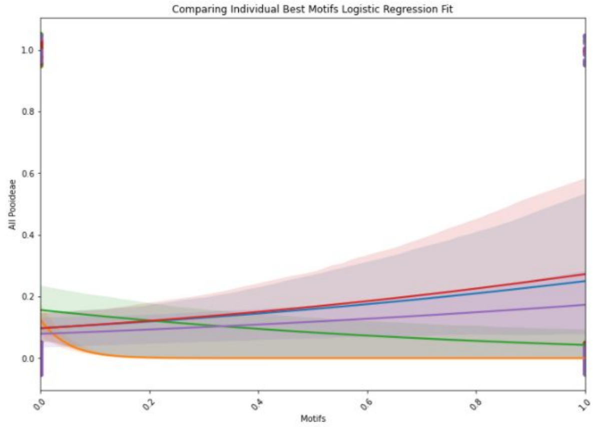
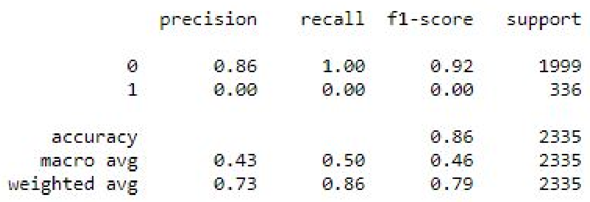


Figure 6: Shows the five motifs individual contributions to the logistic model. Each line represents if just that feature alone was used to construct the final logistic model. There are two trends that the motifs seem to follow. The first trend is that the motifs are important in identifying species not in core *Pooideae* and relatives (green and orange trend lines). The other trend is that the motifs are very important in the identification of core *Pooideae* and relatives (red, blue, purple).

4.3 Accuracy and Tree Reconstruction

The accuracy of the model when used by the test data was determined, 0.86, along with the confusion matrix values (Table 1). The confusion matrix is lacking true and false negatives and instead only has true (0) and false (1) positives. This can be expected since most motifs are present in several genes. The five motifs selected for by the feature selection were then the only motifs plotted on the new genetic visual next to the phylogeny (Figure 7). Some of these motifs are not in the three main cis-regulating regions under review since two of the motifs were useful in not being identified in core *Pooideae* and relatives.

Table 1: Show the accuracy and confusion matrix. The confusion matrix is the first two rows, with the 1999 value representing true positives and the 336 representing false positives. No true negatives or false negatives were found due to the nature of the data and logistic modeling. The accuracy from the test data is shown to be 0.86 when modeled on the features determined from the training data.


```
              precision    recall  f1-score   support

           0       0.86      1.00      0.92      1999
           1       0.00      0.00      0.00       336

    accuracy                           0.86      2335
   macro avg       0.43      0.50      0.46      2335
weighted avg       0.73      0.86      0.79      2335
```

5. DISCUSSION

With the logistic regression model constructed returning a testing data accuracy of 0.86 the model seemed to be relatively representative (Table 1). The motifs selected for by feature selection all have strong statistically significant p-values which means these motifs are a good fit for predicting (Figure 4, Figure 6). Thus, the motifs selected to be most important in identifying core *Pooideae* species and relatives can be thought of as good indicators for the evolution of this clade. Additionally, the motifs were shown to not always be in cis-regulating regions for other species which supports that these motifs may be influential in the evolution of core *Pooideae* and relatives (Figure 7). However, since some of the motifs identified are thought to be good at identifying species not being in core *Pooideae* and relatives these motifs may not be useful in the context of understanding the evolution of vernalization across all grasses (Figure 6). The two motifs (orange trend line: CACAGTATGGAC, green trendline: CTCTCTCTCTC) that represent motifs not being in the core

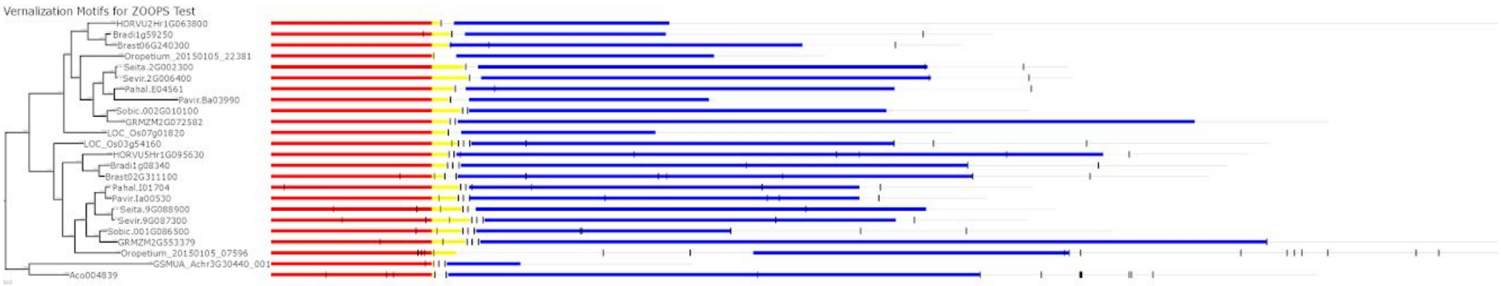


Figure 7: Is the same phylogeny as in Figure 3, except the only motifs shown are the five used in the logistic regression model.

Pooideae and relatives can be used to construct mutants still (Figure 6). However, these motifs can not be mutated within the core *Pooideae* and relatives since they are not present in all of them. Instead these two motifs would have to be mutated in whatever species is used as the outgroup, or not in core *Pooideae* and relatives, when doing a mutant experiment assessing the regulation contribution of each of those two motifs. This is important to acknowledge since not all cereal crops are within the core *Pooideae* and relatives. It would be more beneficial to use the motifs that were found to be important in core *Pooideae* and relatives, because these motifs can be searched for in other species to see if there is a time these motifs evolved within a larger clade of grasses. Overall, the results found are novel and after searching if these motifs have already been identified or not could reveal their importance in regulating and responding to vernalization. Once motifs have been identified as novel or associated with a protein or function will help determine the precedents of which motifs should be experimented with first.

The limitations of the research must also be considered. Since a maximum motif length of 12 was chosen there could be motifs larger in size that were missed. However, if a motif of longer length was heavily conserved then part of the sequence likely would have been selected as a motif by MEME. Another limitation is that motifs found from the feature selection were mainly in FUL1 (Figure 5). This means that this model and the motifs identified for importance may not be the most representative of FUL2 motifs. FUL1 and FUL2 are both important in the vernalization pathway, so being able to understand how both individual genes are being regulated matters for the phenotype. Lastly, an assumption made is that the most important motifs are in cis-regulating regions [15]. Regulation very well could be occurring in trans-regulating regions, but trans regions can exist anywhere in the genome that is not within the gene region. Thus narrowing where to look throughout the entire genome is not feasible. Motifs that occur in cis-regulating regions in one species but not in another also could mean that different species of grasses are having variation within which transcription factors or motifs are actually regulating these genes. This is not far fetched given the thousands of years evolution is bound to have introduced new variation within each grass species genome. However, since there is a close relationship among the grasses clade it seems more likely that the regulation method would be relatively highly conserved [17].

When comparing the results to prior work it is hard to relate because no studies have reviewed vernalization motifs. However, no CAr-G box motif was selected from the logistic model (Figure 4). This was expected given that the CAr-G box motif has been found in all vernalization genes and thus would not be a good motif at identifying just a subset of species like core *Pooideae* and relatives [16, 24]. The genetic diagram constructed can be assumed to be correct as well when compared to the several constructed by prior research (Figure 3, Figure 7) [17]. Unfortunately, time ran out before each motif could be searched in a database to determine its novelty or function. Prior work did do this analysis and results showed wide ranges in motif associated functions and responses [4, 17]. This analysis would be beneficial in narrowing down which motifs are influential. Overall the work done is a good foundational start to a longer project that could result in a better understanding of what genetic attributes are controlling the vernalization pathway. Once the regulation of vernalization, and hence vernalization genes, is understood then possible GMOs or enhanced agriculture techniques could be implemented to make for a more sustainable future in food production as climate change continually tests humanity.

6. FUTURE WORK

To improve the results from the work already done a few corrections and additional processes should be performed. First to revise is the feature selection process. A chi-square approach was good, but incorporating a feature permutation and comparing the accuracy across selection tests would be beneficial. Feature permutation was not performed simply because it was forgotten about when in the thick of the project. However, a permutation test would likely improve the accuracy or help refine which motifs are good characterizers. Next, the model showed there to be only one motif found in FUL2. Thus it would likely be best to rerun the feature selection and model thrice, once for FUL1 motifs, once for FUL2 motifs and lastly once for both genes as was done above. This would allow for more precise motifs for each gene to be identified as defining characteristics of the core *Pooideae* genes. Since these genes, FUL1 and FUL2, are distantly related, paralogs, there may be some overlap in the motifs identified as important, but that is not guaranteed. To widen the range of possibilities separate tests should be done.

The motifs identified by feature selection to be important in determining core *Pooideae* and relatives should be searched for in a motif database. A good database containing information on already discovered plant motifs is PLACE [10]. Besides just searching for the motifs identified by the model all the motifs found to be in cis-regulating regions should be searched in the database. All motifs searched for in the database should have their associated function or proteins recorded, even if there is no associated function. No associated function would mean the motif is novel and that there is no prior research done on what that motif's function is. Novel motifs are important because they can have effects not yet understood. Additionally, motifs known to be associated with regulating proteins should be considered important as well. Any motif that is novel or identified with a transcription factor, regulating protein, should thus be used as possible sites for mutation in a future experiment to test the effect that each individual motif has on the phenotypic response to vernalization.

Lastly, a full separate experiment should be conducted. The experiment should generate mutants from the most important motifs thought to be involved in regulation. The mutant plants should be subject to different environments, one with a longer winter, one with almost no winter and a control group where the winter reflects the normal overwintering time. Each mutant for each treatment should have values measuring the variation in phenotypic response, days to flower and yield of flower, to vernalization. This hopefully can expose which motifs are essential in this pathway and important for further research into making successful mutants that need to endure less or no wintering.

7. CONCUSSION

This project determined several motifs throughout representative species of the grass family. The motifs identified were in genes relating to vernalization and then further selected for motifs residing in cis-regulating regions. The most important motifs for identifying core *Pooideae* species and relatives were determined. All of the motifs identified in cis-regulating regions may be regulating these vernalization genes, however should be further examined with a comparison to motifs whose function have already been determined. The most important motifs identified can be used in a future experiment looking at the evolution of core *Pooideae* species response to vernalization which could give relatable results to the rest of *Poaceae*. Understanding how vernalization is being regulated is dire in terms of food security with the warming climate. Sooner or later the winters will be minimal in areas where cereal crops currently thrive. To combat the risk of famine the best way to head into the future with a warming climate is to try to indepthly understand the processes contributing to crop growth and sustainable agriculture. The survival of humanity depends on the cohesiveness of the planet to remain intact, which can not be easily understood with a large amount of food scarcity crippling society. This exploratory experiment is an urgent call for the importance of food security, something not well considered, when discussing the risks and effects of climate change.

8. ACKNOWLEDGMENTS

Dr. Jill Preston provided me with a great background in this field of research. Most of the genetic results were confirmed by her and it has been a pleasure to learn from her. Additionally, I would like to thank my professor Dr. Nick Cheney for providing all the computational wisdom. He was also helpful in guiding the project into the right direction for building the model. Finally, the outside platforms used were super and it is wonderful that they are free and readily available for public use (Phytozome, MEME, MAFFT and MrBayes).

9. REFERENCES

[1] Alexandre, C., M., & Hennig, L., (2008). *FLC* or not *FLC*: the other side of vernalization, *Journal of Experimental Botany*, Volume 59, Issue 6, April 2008, Pages 1127–1135, https://doi.org/10.1093/jxb/ern070

[2] Bailey, T., L., Bodén, M., Buske, F., A., Frith, M., Grant, C., E., Clementi, L., Ren, J., Li, W., W., Noble, W., S., (2009). "MEME SUITE: tools for motif discovery and searching", *Nucleic Acids Research*, 37:W202-W208, 2009. [full text]

[3] Chen, A., Dubcovsky, J., (2012). Wheat TILLING Mutants Show That the Vernalization Gene *VRN1* Down-Regulates the Flowering Repressor *VRN2* in Leaves but Is Not Essential for Flowering. PLOS Genetics 8(12): e1003134. https://doi.org/10.1371/journal.pgen.1003134

[4] Chern, D., Wong, J., Gutierrez, R. L., Alan, G., & Castellarin, S. D. (2017). Genome-wide analysis of cis-regulatory element structure and discovery of motif-driven gene co-expression networks in grapevine. 24(January), 311–326. https://doi.org/10.1093/dnares/dsw061

[5] Cock PA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B and de Hoon MJL (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25, 1422-1423. http://dx.doi.org/10.1093/bioinformatics/btp163

[6] Deng, W., Casao, M. C., Wang, P., Sato, K., Hayes, P. M., Finnegan, E. J., & Trevaskis, B. (2015). and other key traits of cereal crops. (May 2014). https://doi.org/10.1038/ncomms6882

[7] Dixon, L. E., Karsai, I., Kiss, T., Adamski, N. M., Liu, Z., Ding, Y., … Griffiths, S. (2019). VERNALIZATION1 controls developmental responses of winter wheat under high ambient temperatures. https://doi.org/10.1242/dev.172684

[8] Dubcovsky J, Yan L. (2003). Allelic variation in the promoter of *Ap1*, the candidate gene for *Vrn-1*. In: Pogna N (ed) Proceedings of the 10th international wheat genetics symposium, vol 1. Paestum, Italy, pp 243–246

[9] Fu, D., Szűcs, P., Yan, L. *et al.* Large deletions within the first intron in *VRN-1* are associated with spring growth habit in barley and wheat. *Mol Genet Genomics* 273, 54–65 (2005). https://doi.org/10.1007/s00438-004-1095-4

[10] Higo, K., Ugawa, Y., Iwamoto, M. and Korenaga, T. "Plant cis-acting regulatory DNA elements (PLACE) database: 1999" Nucleic Acids Research, Volume 27, Issue 1, 1999, Pages 297-300.

[11] Huelsenbeck, J.P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17:754-755.

[12] Huerta-Cepas, J., Serra, F., Bork, P., (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data, *Molecular Biology and Evolution*, Volume 33, Issue 6, June 2016, Pages 1635–1638, https://doi.org/10.1093/molbev/msw046

[13] Katoh, K., Misawa,K., Kuma, K., & Miyata, T., (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Research*, Volume 30, Issue 14, 15 July 2002, Pages 3059–3066, https://doi.org/10.1093/nar/gkf436

[14] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, Takashi Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Research*, Volume 30, Issue 14, 15 July 2002, Pages 3059–3066, https://doi.org/10.1093/nar/gkf436

[15] Li, J., Yuan, J., & Li, M. (2014). Characterization of Putative cis -Regulatory Elements in Genes Preferentially Expressed in Arabidopsis Male Meiocytes. 2014. http://dx.doi.org/10.1155/2014/708364

[16] Pidal, B., Yan, L., Fu, D., Zhang, F., Tranquilli, G., Dubcovsky, J., (2009). The CArG-Box Located Upstream from the Transcriptional Start of Wheat Vernalization Gene *VRN1* Is Not Necessary for the Vernalization Response, *Journal of Heredity*, Volume 100, Issue 3, May-June 2009, Pages 355–364, https://doi.org/10.1093/jhered/esp002

[17] Powell, R. V, Willett, C. R., Goertzen, L. R., & Rashotte, A. M. (2019). Lineage specific conservation of cis- regulatory elements in Cytokinin Response Factors. Scientific Reports, (February), 1–11. https://doi.org/10.1038/s41598-019-49741-6

[18] Python Software Foundation. Python Language Reference, version 3.8. Available at http://www.python.org

[19] Ronquist, F., and J.P. Huelsenbeck. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

[20] Ronquist, F., M. Teslenko, P. van der Mark, D.L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M.A. Suchard, and J.P. Huelsenbeck. 2012. MRBAYES 3.2: Efficient Bayesian phylogenetic inference and model selection across a large model space. Syst. Biol. 61:539-542.

[21] Shi, C., Zhao, L., Zhang, X., Lv, G., Pan, Y., & Chen, F. (2019). Gene regulatory network and abundant genetic variation play critical roles in heading stage of polyploidy wheat. 1–16. https://doi.org/10.1186/s12870-018-1591-z

[22] Stewart AJ, Hannenhalli S, Plotkin JB. Why transcription factor binding sites are ten nucleotides long. Genetics. 2012 Nov;192(3):973-85. doi: https://doi.org/10.1534/genetics

[23] Woods, D., P., McKeown, M., A., Dong, Y., Preston, J., C., Amasino, R., M., (2016). Evolution of *VRN2/Ghd7*-Like Genes in Vernalization-Mediated Repression of Grass Flowering. Plant Physiology Apr 2016, 170 (4) 2124-2135; https://doi.org/10.1104/pp.15.01279

[24] Yan L, Loukoianov A, Tranquilli G, Helguera M, Fahima T, Dubcovsky J. Positional cloning of wheat vernalization gene *VRN1*, *Proc Natl Acad Sci U S A*, 2003, vol. 100 (pg. 6263-6268) https://doi.org/10.1073/pnas.0937399100

10. About the authors:

Isaac Racine is a senior at the University of Vermont majoring in Biological Sciences and minoring in Statistics. He has suitable domain knowledge of the topic but is hoping to improve his data science and coding capabilities by continuing his education in the coming years.