ANALYSIS AND COMPARISON OF REGULATORY REGIONS BETWEEN *ORYZA SATIVA* AND NON-PLANT EUKARYOTIC SPECIES

by

Shiva Rawat

A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Master of Science
Bioinformatics and Computational Biology

Committee:

_____	Dr. Ancha Baranova, Thesis Director
_____	Dr. Don Seto, Committee Member
_____	Dr. Tatiana Tatarinova, Committee Member
_____	Dr. Iosif Vaisman, Director, School of Systems Biology
_____	Dr. Donna Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
_____	Dr. Peggy Agouris, Dean, College of Science
Date: _____	Spring Semester 2019 George Mason University Fairfax, VA

Analysis and Comparison of Regulatory Regions Between *Oryza Sativa* and Non-Plant Eukaryotic Species

A thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science at George Mason University

By

Shiva Rawat
Bachelor of Arts
University of Virginia, 2013

Director: Ancha Baranova, Professor
School of Systems Biology

Spring Semester 2019
George Mason University
Fairfax, VA

ACKNOWLEDGEMENTS

I would like to thank my loving parents and George Mason University for providing the support and resources necessary to accomplish this endeavor. I would like to thank Professor Baranova for offering this research opportunity and advising me throughout this process. I would also like to thank Professor Tatarinova for providing the data required for analysis and Professor Smirnov for providing guidance during this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

ANALYSIS AND COMPARISON OF REGULATORY REGIONS BETWEEN *ORYZA SATIVA* AND NON-PLANT EUKARYOTIC SPECIES

Shiva Rawat, MS

George Mason University, 2018

Dissertation Director: Ancha Baranova, Professor

Deciphering the non-coding regions of the eukaryotic genes continues to be an important transdisciplinary area spanning bioinformatics and experimental biology fields. Upstream of a eukaryotic gene there is a region of regulatory DNA sequences, which serve as genetic switches for modulating its expression. Various computational methods have been implemented to locate specific functionally important genomic elements. Nonetheless, accurately detecting promoters, transcription start sites (TSS), and transcription factor binding sites (TFBS) is a conundrum as even the finest and most sophisticated computational tools produce a significant amount of false positive and false negative predictions. In this study, our objective was to accurately classify regulatory regions in several eukaryotic species and elucidate their roles. Information of several transcription factors (TFs) of rice (*O. sativa*), their corresponding binding sites and the distribution of these sites near TSSs of rice genome were obtained from (Triska, et al., 2017), (Tatarinova, et al., 2016). For each transcription factor, we determined its orthologues in other species and compared their binding sites to those of *O. sativa* transcription factors. After mapping

the identified TFs to non-plant eukaryote genomes, we discovered their putative binding sites in non-plant eukaryote species, and analyzed the distribution of distances between corresponding TFBS and TSS. We have shown that the genome distributions of orthologous TFBSs in rice (*O. sativa)*, fruit fly (*D. melanogaster)*, two species of rodents (*M. musculus* and *R. norvegicus*), as well as humans (*H. sapiens*) share similar characteristics, thus, proving a conservative nature of TF-TFBS interactions at a genomic scale.

INTRODUCTION

The genomes of eukaryotic organisms are larger and more intricate than those of prokaryotic organisms because eukaryotic chromosomes contain greater amount of noncoding DNA (Cooper, 2000). Once believed to be mostly junk and useless, noncoding DNA is now considered to harbor massive amounts of information necessary for the activation and deactivation of genes.

The areas of noncoding DNA that control the expression of genes are referred to as *regulatory regions* and contain a variety of specific nucleotide patterns known as *sequence motifs*. These motifs act as environmental sensors, stochastic signal generators (as for the c-MYC gene) (Halazonetis, et al., 1991) and as binding sites for transcription factors (a specialized set of DNA-binding proteins that regulate gene expression by attaching itself to appropriate transcription factor binding sites). These binding sites are often referred to as response elements or cis-regulatory elements. Promoters, enhancers and silencers are specific types of regulatory sequences that contain assortment of TBFSs. A promoter is a specific region of DNA that initiates transcription by guiding RNA polymerase and transcription factors to DNA binding sites for assembly into a preinitiation complex (Alberts, et al., 2002). Enhancers are regulatory sequences that serve as DNA binding sites for transcription factors known as activators to increase the rate of transcription. In contrast, silencers are regulatory sequences that act as DNA binding sites for transcription factors known as repressors, which decrease the rate of transcription from adjacent locus. Each promoter regulates transcription based on the cell/tissue

type, developmental stage, intra/intercellular signals, and environmental conditions (Shahmuradov, et al., 2017). The interaction of countless DNA-binding and bending proteins, which commonly form one or another multi-subunit complex, and their binding sites make eukaryotic transcriptional regulation an extremely convoluted process (Eckardt, 2014).

Transcription start sites (TSS) and core promoters are essential regulatory regions that are often incorrectly identified with current computational methodologies. A core promoter is the minimal portion of the promoter region required to begin transcription; it contains the TSS along with TFBSs (Shahmuradov, et al., 2017). Studies on mammalian and plant genomes have revealed that many eukaryotic genes are associated with multiple distinct promoters. Moreover, eukaryotic promoters are typically characterized by multiple TSSs, and can be classified based on the distribution and utilization of their set of TSSs. Consequently, the association with several distinct promoters allows for a single gene to encode various protein isoforms (Sandelin, et al., 2007).

TATA-box is a highly conserved regulatory sequence motif contained within the core promoter of eukaryotes, frequently located 30 nucleotides upstream of TSS. Its ancient origin can be derived from its appearance in 30-60% of all eukaryotic promoters, including those from the organisms as diverse as yeast, plants, and metazoans (Tatarinova, et al., 2009). Genes that possess the TATA-box are typically expressed under conditions of stress, are tissue-specific, and demonstrate high plasticity by tolerating extremes in expression levels (Troukhan, et al., 2009).
On the other hand, large-scale genome studies on various eukaryotes, including plants, have revealed promoters that lack the TATA-box and rely on other non-coding DNA sequences for the initiation of transcription. These TATA-less promoters are typically associated with housekeeping genes and rely upon short sequence motifs such as the Initiator element (INR), BRE (B recognition element for TFIIB), DPE (downstream promoter element), MTE (motif ten

element), TCT (polypyrimidine initiator), and Sp1 (specificity protein 1) (Troukhan, et al., 2009). These sequence motifs also belong to the set of core promoter elements and help in regulating the initiation of transcription.

Transcription factors play a major role in initiating transcription through protein-protein interactions with the general transcription machinery, which is composed of RNA polymerase and various transcription factors and known as the preinitiation complex. Additionally, transcription factors interact with specific DNA sequences known as enhancers and silencers to activate and suppress gene expression. Enhancers and silencers are typically located close to the promoter region and are represented by exceptionally selective binding sites that require a specific DNA recognition process involving non-covalent interactions between exposed surfaces of the DNA molecule and the properly oriented structural motifs of a given transcription factor (Raab and Kamakaka, 2010). In fact, enhancers and silencers might have evolved as specialized derivatives of promoters (Raab and Kamakaka, 2010). Deciphering the mechanisms behind the various interactions that occur among transcription factors and their binding sites is essential to elucidating the complex process of transcriptional regulation.

Due to the complex architecture of eukaryotic promoters, computational analysis of these regulatory regions is challenging and requires sophisticated algorithms (Singh, et al., 2015). Some regulatory sequence motifs such as the TATA-box or the CA-motif have been extensively analyzed both experimentally and computationally. However, more comprehensive analyses of regulatory motif architecture integrated with *in silico* modelling may allow for superior prediction power (Venter, et al., 2009). Oligonucleotide content-based neural network and linear discriminant approaches are examples of novel complex algorithms that have been utilized by many TSS prediction software tools. These tools have been successful in predicting TATA-containing promoters with high specificities and sensitivities for the model plant *Arabidopsis*

thaliana (Tatarinova, et al., 2013). However, the TSS prediction accuracies of these computational tools are low for genomes with more complex designs such as *H. sapiens* and *O. sativa*. In fact, the quality of promoter predictions declines with an increase in genome complexity because of the presence of alternative TSSs, and commonly results in erroneous promoter calculations that can be observed through large false positive and false negative error rates. The challenge of identifying TSSs at near 100% accuracy has eluded the most advanced and successful techniques of promoter mapping such as full-length cDNA analysis.

Alternative methods of eukaryotic promoter prediction involve the use of previously discovered and experimentally characterized TFBSs. Two decades ago, Kondrakhin and Kel employed a technique that improved promoter prediction accuracy by combining position weight matrices of TFBSs with the detection of TATA-boxes (Kondrakhin, et al., 1995). Later, the method developed by Prestridge generated low false positive rates and involved the use of a weighted TATA matrix in conjunction with the scoring profile of every TFBS (Prestridge, 1995). Nonetheless, the lack of effective tools to predict individual TSSs along with the dearth of species-specific TFBS models for training collectively limit the utility of these two methods.

The approach invented by Troukhan is a deterministic method that has displayed incredible accuracy for *Arabidopsis thaliana* promoters (Troukhan, et al., 2009). Initially, its efficiency was limited to less complex genomes due to its prediction of a single promoter and TSS per gene. Subsequently, the approach was upgraded to the NPEST algorithm in order to analyze the architecture of more complex eukaryotic promoters and their collection of TSSs (Tatarinova, et al., 2013).

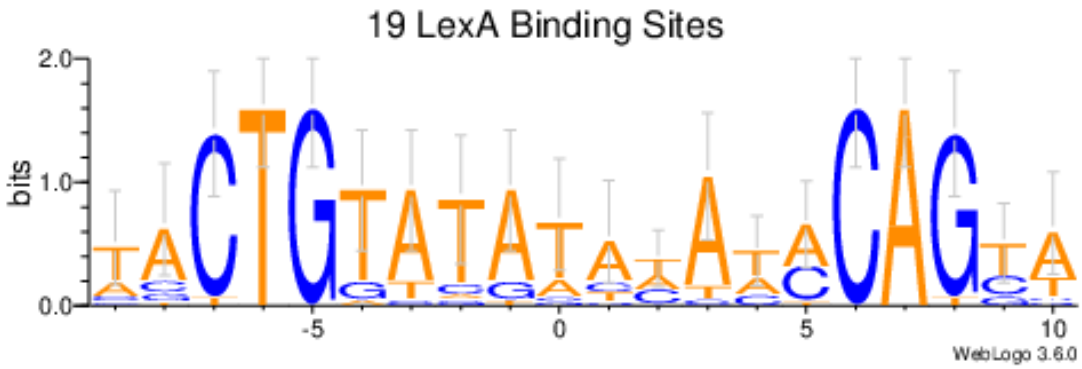
NPEST utilizes a non-parametric maximum likelihood approach for the prediction of alternate TSSs and produces results superior to that of other algorithms such as TAIR when run on databases containing experimentally confirmed promoters.

Position Weight Matrices (aka PWM) are commonly utilized for discovery and the systematic analysis of sequence motifs embed in nucleotide and protein sequences. PWMs play a pivotal role in advanced software algorithms that characterize cis-regulatory elements, like PSI-BLAST (Bhagwat et al., 2007). PWM is the most natural and effective ways to represent TFBS. The algorithms that use PWMs use the information about patterns of positional nucleotide frequencies and compare them with an unknown sequence in order to find potential matches and to calculate the statistical significance of matches at each location for a particular binding site (Stepanova, et al., 2005).

Applications that use PWM for motif discovery and analysis take an aligned set of sequences along with their specified background frequencies as the input. Three methods for specifying the background frequencies have been developed already, and are well described in existing scientific literature. Simplest techniques, which involve using identical background frequencies, were developed for characterizing splice sites. (Schwartz, et al ., 2008; Sheth, et al., 2006). A second approach calculates background frequencies from input sequences themselves and a third method is to define the background frequency based on the context of the problem (Xia, 2012). The output generated from a standard PWM analysis consists of the predicted PWM, the site-specific and motif information content, and the position weight matrix scores (PWMS) for each sequence used as input (Xia, 2012).

Sequence logos are graphical representations of regulatory sequence motifs that utilize bits of information to depict nucleotide or amino acid conservation (Schneider, 2002), for example, various binding sites in DNA or functional units in proteins. Additionally, sequence logos are often used to characterize PWM of transcription factors. This is accomplished by plotting the site-specific nucleotide frequency values versus site positions. Consequently, each sequence logo is represented by a stack of letters at every position of the sequence, where the

relative size of each letter reflects its frequency in the sequence and the total height of the stack of letters represents the information content of the position, in bits.



(http://weblogo.threeplusone.com/examples/lexA.png)

Figure 1. Example of a sequence logo for the LexA transcription factor

The figure above (Figt.1) provides an illustration of a sequence logo for the LexA TFBS. For each position, the size of each individual letter represents the nucleotide frequency while the size of the overall stack depicts the information content, measured in bits.

Why we need accurate prediction of the promoters for annotation of agriculturally important genomes

Accurately predicting regulatory regions in *O. sativa* crops has major global implications on agriculture and food security. *O. sativa* is an essential staple crop that provides 20% of daily calories for the world's population (World Rice Statisitics, https://www.irri.org). Moreover, rice serves as a model organism for cereal biology and plant functional genomics research which significantly impacts global food security through crop genetic improvements. Next-generation sequencing (NGS) has provided a wealth of high-quality rice genomic sequences that have

significantly contributed to the growth and evolution of plant functional genomics (Jiang, et al., 2011). Accordingly, a diverse set of functional genomic platforms has been developed for the genetic analysis of rice including full-length cDNA libraries, massive mutant libraries, gene expression microarrays, global expression profiles, RNA sequencing, and a collection of germplasm resources. There are large curated databases which host available information on genomics, transcriptomics, metabolomics, and proteomics of rice.

Although the genome of *O. sativa* has been fully sequenced and is well studied, the complex regulatory networks controlling the expression of genes and the development of phenotypes remain an unsolved mystery. Consequently, elucidating the mechanisms essential to gene expression is vital for the selection of favorable agronomic traits in rice. Yield, grain quality, resistance to disease and pests, nutrient-use efficiency, abiotic stress resistance, and reproductive development are among the agronomic traits targeted for enhancement through genetics biomarker-aided breeding (Li, et al., 2018). Because genomic resources are readily available for rice, and of its relatively small genome size, this plant has been increasingly used as a proving ground for various genome editing technologies. There is a hope that advanced versions of genome editing kits, including CRISPR/Cpf1 system and base editors would be instrumental in accelerating the pace of crop improvement (Mishra, et al., 2018).

SPECIFIC AIMS

PWM-evaluated TF binding to genomic DNA displays marked positional bias in relationship to experimentally mapped TSS. This phenomenon was first described by Weirauch, Yang (Weirauch, et al., 2014), who showed that this observation holds for both plants and animals. Positional clustering of binding sites shows that the rice *O. sativa*, the major staple food crop of the world, utilizes three distinct classes of transcription factors: those that bind preferentially to the [-500,0] region (188 "promoter-specific" transcription factors), those that bind preferentially to the [0,500] region (282 "5' UTR-specific" TFs) [Triska et al., 2017], and those that have no preference for binding to one specific region ("promiscuous transcription factors").

This study aimed to determine if promoter–specific and 5'-UTR specific TFs of *O. sativa* have orthologues in non-plant eukaryotic species. For each discovered orthologous transcription factor, PWMs for their binding sites were compared to those of *O. sativa* and analyzed for the distribution of their positions relative to the nearest TSSs and AUGs in respective genomes. The sequence motifs of the binding sites were compared as well; the latter allowed us to detect possible species-specific base difference at each position.

MATERIALS AND METHODS

Orthologous Transcription Factors

SWISS-MODEL

The SWISS-MODEL (https://swissmodel.expasy.org/) structural bioinformatics web-server was used to search for non-plant eukaryotic transcription factors homologous to *O. sativa* transcription factors. SWISS-MODEL is a server with a fully automated pipeline that specializes in three-dimensional protein structure homology modeling. The protein models are constructed from templates of homologs that are obtained from experimentally solved structures in the Protein Data Bank (PDB). To construct the most accurate protein model, an alignment of the target sequence with the template occurs upon the retrieval of the template.

PDB

The Protein Data Bank (PDB) is a database that contains experimentally solved three-dimensional protein structures (https://www.rcsb.org/). The experiments typically conducted to reveal the three-dimensional structure of proteins are X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy.

UniProt

UniProt (https://www.uniprot.org/) is a repository for protein sequences and their biological function, built by combining Swiss-Prot, TrEMBL, and PIR-PSD databases. The protein sequences have been generated from whole genome sequencing projects while their biological functions have been derived from research literature.

BLAST

BLAST stands for Basic Local Alignment Search Tool. This suite of programs that identify regions of local similarity between DNA and protein sequences. The programs accomplish this task by aligning query sequences against those present in a selected target database. BLAST calculates the statistical significance of matches found in the sequences database. Consequently, BLAST aids in discerning functional and evolutionary relationships between sequences along with determining members of gene families. The various BLAST tools are provided by the National Center for Biotechnology Information (NCBI) and can be accessed from the NCBI BLAST homepage (https://blast.ncbi.nlm.nih.gov/Blast.cgi). In addition to the web version of BLAST, a standalone version can be downloaded from the homepage. This allows the execution of BLAST software on a local computer with downloaded copies of NCBI BLAST databases.

O. sativa Protein Structure Comparison

jFATCAT_flexible

jFATCAT_flexible is a program for comparing protein structures through pairwise structure alignments. The program takes PDB files as input and generates alignment results along with a superimposed image of the two protein structures being compared.

O. sativa Transcription Factors/Bindings Sites

Jaspar

Jaspar (http://jaspar.genereg.net) is an open-access and manually curated TFBS database that includes profiles for eukaryotic organisms within six taxonomic groups (Khan, et al., 2018). The database is non-redundant with transcription factor binding sites stored as position frequency matrices (PFM) and transcription factor flexible models (TFFMs). PFMs contain the nucleotide

frequencies at each position in a set of aligned sequences believed to share a common ancestry as each sequence is bound to the same transcription factor.

The entire collection of PFMs for the *O. sativa* plant species was downloaded and converted into PWMs. PWMs are probabilistic models that have gained favor over consensus sequences because the former is a more effective approach for depicting patterns in biological sequences and discovering DNA-binding motifs. The transformation from PFM to PWM was completed by first dividing the nucleotide frequencies at each position of the matrix by the background letter frequency and then calculating the logarithm of a likelihood ratio for each. The highest values at each position are then added to obtain the greatest probability of a matching sequence.

PlantTFDB

Currently in its fourth version, the PlantTFDB (http://planttfdb.cbi.pku.edu.cn) is a repository of transcription factors from a diversity of plant species. The PlantTFDB houses records of 320,370 transcription factors from 165 species along with a plethora of functional and evolutionary annotation for each protein. Furthermore, the most recent upgrade to the database involved additional annotation in the form of non-redundant transcription factor binding motifs obtained from experiments, an assortment of regulatory elements classified from high-throughput sequencing data, and the regulatory interactions curated from literature (Jin, et al., 2016). A collection of *O. sativa* transcription factors and the corresponding binding motifs was downloaded for comparative analysis with other eukaryotic species from the PlantTFDB.

Eukaryotic Transcription Factors/Binding Sites

TRANSFAC

Transcription factor binding site data was obtained from the TRANSFAC database (http://gene-regulation.com/pub/databases.html). TRANSFAC is a manually curated database

that contains transcription factors and their corresponding binding sites for a variety of eukaryotic organisms. The database is currently organized, updated, and disseminated by geneXplain GmbH, Wolfenbüttel, Germany. Data files containing predicted rice binding sites were generated from the database. Binding sites for flies, mice, rats, and humans were also produced and stored in a massive data file.

Sequence Logos

WebLogo 3

The sequence logos for the consensus sequences were produced by the WebLogo 3 application (http://weblogo.threeplusone.com/). WebLogo 3 is an online tool that takes a list of aligned sequences and generates a graphical representation called a logo. The logo depicts the nucleotide conservation of the sequences and provides more information about a binding site than a consensus sequence. A logo is comprised of a stack of letters for each position of a sequence. The height of each letter indicates the frequency of the nucleotide at that position while the overall height of the stack of letters represents the information content, measured in bits, for that position.

RESULTS

O. sativa transcription factors and their non-plant eukaryotic orthologues

A group of *O. sativa* transcription factors and their non-plant eukaryotic orthologues are described in the Table 1. The first column lists the *O. sativa* motif names while the second column displays the corresponding transcription factor protein that binds to the motif. The third and fourth columns list the gene and orthologous transcription factor protein for the non-plant eukaryote while the final column displays the species name.

Table 1. *O. sativa* transcription factor orthologues in other species

O. sativa motif name	*O. sativa* TF protein name	Orthologous TF gene name	Orthologous TF protein name	Orthologous species
ABI4	Ethylene-response TF ABI4 (C7J2Z1)	PF3D7_1466400	Transcription factor with AP2 domain (Q8IKH2)	*Plasmodium falciparum* (isolate 3D7)
ABI5	bZIP transcription factor ABI5 homolog (Q8RZ35)	CEBPB	CCAAT/enhancer-binding protein beta (P17676)	*Homo sapiens* (Human)
ABI5	bZIP transcription factor ABI5 homolog (Q8RZ35)	Mafg	Transcription factor MafG (O54790)	*Mus musculus* (Mouse)
ALFL5	PHD finger protein ALFIN-LIKE 5 (Q60DW3)	BPTF	Nucleosome-remodeling factor subunit BPTF (Q12830)	*Homo sapiens* (Human)
ALFL5	PHD finger protein ALFIN-LIKE 5	TAF3	Transcription initiation factor TFIID subunit 3	*Homo sapiens* (Human)

	(Q60DW3)		(Q5VWG9)	
ALFL5	PHD finger protein ALFIN-LIKE 5 (Q60DW3)	DIDO1	Death-inducer obliterator 1 (Q9BTC0)	*Homo sapiens* (Human)
ALFL5	PHD finger protein ALFIN-LIKE 5 (Q60DW3)	Dido1	Death-inducer obliterator 1 (Q8C9B9)	*Mus musculus* (Mouse)
ALFL5	PHD finger protein ALFIN-LIKE 5 (Q60DW3)	ING4	Inhibitor of growth protein 4 (Q9UNL4)	*Homo sapiens* (Human)
BZR1	Protein BZR1 homolog 1 (Q7XI96)	Myod1	Myoblast determination protein 1 (P10085)	*Mus musculus* (Mouse)
TRAB1	bZIP transcription factor TRAB1 (Q6ZDF3)	JUND	Transcription factor jun-D (P17535)	*Homo sapiens* (Human)
TRAB1	bZIP transcription factor TRAB1 (Q6ZDF3)	FOSB	Protein fosB (P53539)	*Homo sapiens* (Human)
TRAB1	bZIP transcription factor TRAB1 (Q6ZDF3)	Creb1	Cyclic AMP-responsive element-binding protein 1 (Q01147)	*Mus musculus* (Mouse)
GAMYB	Transcription factor GAMYB (Q0JIC2)	MYBL2	Myb-related protein B (Q03237)	*Gallus gallus* (Chicken)
GAMYB	Transcription factor GAMYB (Q0JIC2)	Myb	Transcriptional activator Myb (P06876)	*Mus musculus* (Mouse)
GAMYB	Transcription factor GAMYB (Q0JIC2)	CDC5L	Cell division cycle 5-like protein (Q99459)	*Homo sapiens* (Human)
ZHD5	Zinc-finger homeodomain protein 5 (Q5VM82)	prd	Segmentation protein paired (P06601)	*Drosophila melanogaster* (Fruit fly)
ZHD5	Zinc-finger homeodomain protein 5 (Q5VM82)	al	Homeobox protein aristaless (Q06453)	*Drosophila melanogaster* (Fruit fly)
ZHD5	Zinc-finger homeodomain protein 5	HMBOX1	Homeobox-containing protein 1 (Q6NT76)	*Homo sapiens* (Human)

	(Q5VM82)			
ZHD5	Zinc-finger homeodomain protein 5 (Q5VM82)	Ubx	Homeotic protein ultrabithorax (P83949)	*Drosophila melanogaster* (Fruit fly)
ZHD5	Zinc-finger homeodomain protein 5 (Q5VM82)	exd	Homeobox protein extradenticle (P40427)	*Drosophila melanogaster* (Fruit fly)
HSFC1B	Heat stress transcription factor C-1b (Q942D6)	Hsf	Heat shock factor protein (P22813)	*Drosophila melanogaster* (Fruit fly)
HSFC1B	Heat stress transcription factor C-1b (Q942D6)	HSF2	Heat shock factor protein 2 (Q03933)	*Homo sapiens* (Human)
GATA19	GATA transcription factor 19 (Q0DNU1)	GATA1	Erythroid transcription factor (P17678)	*Gallus gallus* (Chicken)
GATA19	GATA transcription factor 19 (Q0DNU1)	Gata1	Erythroid transcription factor (P17679)	*Mus musculus* (Mouse)
GATA19	GATA transcription factor 19 (Q0DNU1)	GATA3	Trans-acting T-cell-specific transcription factor GATA-3 (P23771)	*Homo sapiens* (Human)
HSFA6B	Heat stress transcription factor A-6a (Q657C0)	HSF1	Heat shock factor protein 1 (Q00613)	*Homo sapiens* (Human)
NFYB2	Nuclear transcription factor Y subunit B-2 (Q5QMG3)	NFYB	Nuclear transcription factor Y subunit beta (P25208)	*Homo sapiens* (Human)
NFYB2	Nuclear transcription factor Y subunit B-2 (Q5QMG3)	Chrac-14	Chromatin accessibility complex 14kD protein (Q9V444)	*Drosophila melanogaster* (Fruit fly)
ALFL5	PHD finger protein ALFIN-LIKE 5 (Q60DW3)	pps	Protein partner of snf, isoform A (Q9VG78)	*Drosophila melanogaster* (Fruit fly)
ZHD4	Zinc-finger	en	Segmentation	*Drosophila*

	homeodomain protein 4 (Q53N87)		polarity homeobox protein engrailed (P02836)	*melanogaster* (Fruit fly)
YAB3	Protein YABBY 3 (Q8L556)	Tox	Thymocyte selection-associated high mobility group box protein TOX (Q66JW3)	*Mus musculus* (Mouse)
YAB3	Protein YABBY 3 (Q8L556)	TFAM	Transcription factor A, mitochondrial (Q00059)	*Homo sapiens* (Human)
YAB3	Protein YABBY 3 (Q8L556)	SOX17	Transcription factor SOX-17 (Q9H6I2)	*Homo sapiens* (Human)
YAB3	Protein YABBY 3 (Q8L556)	HMGB3	High mobility group protein B3 (O15347)	*Homo sapiens* (Human)
OsHAP2I	CCAAT-binding transcription factor subunit B family protein, expressed (Q338K5)	NFYA	Nuclear transcription factor Y subunit alpha (P23511)	*Homo sapiens* (Human)
LOC_Os10g22950	Anther ethylene-upregulated protein ER1, putative, expressed (Q339A6)	TP53BP2	Apoptosis-stimulating of p53 protein 2 (Q13625)	*Homo sapiens* (Human)
LOC_Os10g22950	Anther ethylene-upregulated protein ER1, putative, expressed (Q339A6)	Espn	Espin (Q9ET47)	*Mus musculus* (Mouse)
MADS3	MADS-box transcription factor 3 (Q40704)	Mef2c	Myocyte-specific enhancer factor 2C (Q8CFN5)	*Mus musculus* (Mouse)
MADS3	MADS-box transcription factor 3 (Q40704)	MEF2B	Myocyte-specific enhancer factor 2B (Q02080)	*Homo sapiens* (Human)
MADS3	MADS-box transcription factor 3 (Q40704)	SRF	Serum response factor (P11831)	*Homo sapiens* (Human)

```
Align 5F28.A.pdb 70 with 6C9L.A.pdb 84
Twists 0 ini-len 64 ini-rmsd 0.74 opt-equ 70 opt-rmsd 0.78 chain-rmsd 0.74 Score 189.88 align-len 70 gaps 0 (0.00%)
P-value 1.11e-16 Afp-num 1558 Identity 87.14% Similarity 95.71%
Block  0 afp  8 score 189.88 rmsd  0.74 gap 0 (0.00%)

                 .    :    .    :    .    :    .    :    .    :    .    :    .    :
Chain 1:   21 FTKRKFGLMKKAYELSVLCDCEIALIIFNSTNKLFQYASTDMDKVLLKYTEYNEPHESRTNSDIVEALNK
              1111111111111111111111111111111111111111111111111111111111111111111111
Chain 2:   21 FTKRKFGLMKKAYELSVLCDCEIALIIFNSANRLFQYASTDMDRVLLKYTEYSEPHESRTNTDILETLKR

Note: positions are from PDB; the numbers between alignments are block index
```

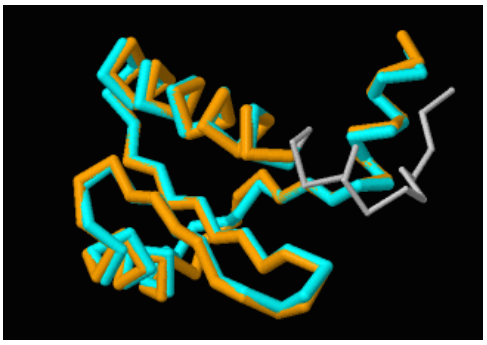


Figure 2. Protein structure alignment results and visualization of the *O. sativa* MADS3 protein and its MEF2 orthologs in human and mice courtesy of the jFATCAT_flexible program.

The jFATCAT_flexible application was utilized to compare the protein structures of the *O. sativa* MADS3 protein and the human/mouse MEF2 protein. At Fig.2, The top image displays the structure alignments along with the p-value, sequence identity, and sequence similarity between the two TFs. The bottom image displays the three-dimensional structural alignment between the MADS3 protein of *O. sativa* and the MEF2 protein in mammals.

A comparison of the transcription factor protein sequences between *O. sativa* and its non-plant eukaryotic orthologues is presented in the Table 2 below. The first column lists the name of the orthologous transcription factor while the next three columns display the comparisons

between the protein sequences in the form of sequence identity, sequence similarity, and coverage. The final two columns list the experimental method for defining the orthologous protein template and the resolution for the method.

Table 2. A comparison of *O. sativa* transcription factor sequences with their orthologues.

Orthologous TF protein name	Sequence identity	Sequence similarity	Coverage	Method	Resolution
Transcription Factor with AP2 domain (Q8IKH2)	19.608	0.275	0.19	X-ray	2.2
CCAAT/enhancer-binding protein beta (P17676)	27.119	0.311	0.152	X-ray	1.85
Transcription factor MafG (O54790)	28.571	0.327	0.144	X-ray	2.8
Nucleosome-remodeling factor subunit BPTF (Q12830)	39.286	0.418	0.217	NMR	NA
Transcription initiation factor TFIID subunit 3 (Q5VWG9)	37.5	0.396	0.217	X-ray	1.703
Death-inducer obliterator 1 (Q9BTC0)	38.889	0.415	0.209	X-ray	1.35
Death-inducer obliterator 1 (Q8C9B9)	51.351	0.497	0.143	NA	NMR
Inhibitor of growth protein 4 (Q9UNL4)	29.63	0.374	0.209	NA	NMR
Myoblast determination protein 1 (P10085)	25.58	0.313	0.144	X-ray	2.8
Transcription factor jun-D (P17535)	27.869	0.329	0.163	X-ray	2.498
Protein fosB (P53539)	26.984	0.334	0.168	X-ray	2.498
Cyclic AMP-responsive element-binding protein 1 (Q01147)	37.5	0.381	0.128	X-ray	3
Myb-related protein B (Q03237)	50.98	0.465	0.184	NMR	NA
Transcriptional activator Myb (P06876)	49.515	0.457	0.186	X-ray	1.68
Cell division cycle 5-like protein (Q99459)	37.209	0.397	0.156	EM	3.6
Segmentation protein paired (P06601)	25.397	0.351	0.272	X-ray	2

Homeobox protein aristaless (Q06453)	25.806	0.352	0.267	X-ray	2.7
Homeobox-containing protein 1 (Q6NT76)	23.438	0.311	0.276	X-ray	2.9
Homeotic protein ultrabithorax (P83949)	20	0.306	0.28	X-ray	2.36
Homeobox protein extradenticle (P40427)	16.923	0.306	0.28	X-ray	2.36
Heat shock factor protein (P22813)	43.617	0.408	0.376	NMR	NA
Heat shock factor protein 2 (Q03933)	51.724	0.433	0.348	X-ray	1.728
Erythroid transcription factor (P17678)	41.463	0.39	0.151	NMR	NA
Erythroid transcription factor (P17679)	46.154	0.412	0.144	X-ray	1.98
Trans-acting T-cell-specific transcription factor GATA-3 (P23771)	43.902	0.395	0.151	X-ray	2.8
Heat shock factor protein 1 (Q00613)	47.222	0.414	0.269	X-ray	2.91
Nuclear transcription factor Y subunit beta (P25208)	73.118	0.53	0.522	X-ray	3.08
Chromatin accessibility complex 14kD protein (Q9V444)	35.354	0.369	0.556	X-ray	2.4
Protein partner of snf, isoform A (Q9VG78)	37.037	0.42	0.209	X-ray	1.4
Segmentation polarity homeobox protein engrailed (P02836)	29.825	0.361	0.137	X-ray	2.1
Thymocyte selection-associated high mobility group box protein TOX (Q66JW3)	31.111	0.351	0.144	NMR	NA
Transcription factor A, mitochondrial (Q00059)	34.146	0.377	0.131	X-ray	2.5
Transcription factor SOX-17 (Q9H6I2)	29.268	0.358	0.131	X-ray	2.4
High mobility group protein B3 (O15347)	26.829	0.34	0.131	NA	NMR
Nuclear transcription factor Y subunit alpha (P23511)	45.714	0.416	0.422	X-ray	3.08
Apoptosis-stimulating of p53 protein 2 (Q13625)	34.579	0.354	0.105	X-ray	2.2
Espin (Q9ET47)	30	0.324	0.108	X-ray	1.65
Myocyte-specific enhancer	50	0.438	0.373	X-ray	2.9

factor 2C (Q8CFN5)					
Myocyte-specific enhancer factor 2B (Q02080)	52.632	0.445	0.322	X-ray	2.3
Serum response factor (P11831)	38.571	0.39	0.297	X-ray	3.15

O. sativa Data Extraction

The regulatory regions of the *O. sativa* genome are comprised of millions of unique binding sites that are widespread throughout the genome. The nomenclature for a collection of binding sites is based on the name of the gene they regulate or the transcription factor that binds to each of the sites. Sequence motifs that characterize each binding site are typically composed of anywhere between 5-15 DNA nucleotides in length. Millions of *O. sativa* TFBSs were collected in a text file *rice_promoter_Sites.bed*, which is 1.18 GB in size. This file was compiled of records stored in the TRANSFAC database. It contains over forty million lines of data, with each line representing a particular *O. sativa* TFBS located within a 2000 nucleotide long sequence (TSS+1000, TSS-1000).

Several types of data were present within each line of the file, including the name of each sequence motif, the chromosome number and mRNA transcript of each binding site, and the position of each binding site relative to the start of the mRNA fragment. An average distance of the binding site positions to the nearest TSS was calculated for each individual motif through implementing a script coded in the Perl programming language.

Program 1: TFBS_extraction.pl

When executed, this script calculates average distance from nucleotide binding site position to the nearest TSS for a given data file and sequence motif. To run this program, the file name with the corresponding binding site data and the specific motif being searched for are

required as inputs. The beginning and end position of each binding site relative to the start of a corresponding 2000 nucleotide long mRNA fragment (TSS+1000, TSS-1000) are listed in the file. The nearest TSS for each binding site is located 1000 nucleotides from the start of each 2000 nucleotide long mRNA fragment. Thus, in order to calculate the binding site position relative to the nearest TSS, 1000 is subtracted from the binding site location. Once all binding site positions for each sequence motif have been computed, the average is taken.

Table 3 (below) contains the data describing the average distance to the nearest TSS for an entire collection of *O. sativa* motifs. The binding site names follow the TRANSFAC notation, with each name containing a unique identifier. The first letter represents the species name, the next group of letters before the underscore provides a description of the transcription factor that binds to particular binding site, and the numbers following the underscore refer to the specific type of the binding site associated with given transcription factor. Each motif name is associated with thousands of TFBSs whose average distance to the nearest TSS is calculated. The first column lists the motif names, while the second column presents the average distance from binding site position to the nearest TSS.

Table 3. Average distances to the nearest TSS for binding sites recognized by various transcription factors of *O. sativa.*

*P-values reflect probabilities that TFBS for given transcription factor accumulated within 200nt of a promoter by chance.

***O. sativa* motif name**	**Number of binding sites**	**Average binding site distance relative to nearest TSS**	**Standard deviation for average binding site distance (+/-)**	**P-value**
P$ABI3_01	20605	524	286	2.20E-16
P$AHL20_02	9083	539	269	2.20E-16
P$AHL25_01	18268	544	269	2.20E-16
P$ALFIN1_Q2	12418	353	278	2.20E-16

P$ARF8_01	44420	500	281	2.20E-16
P$ARR1_01	143447	522	282	2.20E-16
P$ARR10_01	71154	528	281	2.20E-16
P$ASR1_01	379224	441	295	2.20E-16
P$AT1G21910_01	19139	397	269	2.20E-16
P$AT1G26590_01	289055	563	272	2.20E-16
P$AT1G26610_01	140901	544	279	2.20E-16
P$AT1G49120_01	13819	383	260	2.20E-16
P$AT1G53910_01	86303	360	266	2.20E-16
P$AT1G67970_01	110238	502	292	2.20E-16
P$AT1G68550_01	116820	365	266	2.20E-16
P$AT2G15660_01	1674	549	279	2.20E-16
P$AT2G38090_01	936	518	278	2.20E-16
P$AT2G41690_01	96708	460	284	2.20E-16
P$AT2G47520_01	89192	361	266	2.20E-16
P$AT3G18650_01	907	566	273	2.20E-16
P$AT3G25890_01	10717	424	275	2.20E-16
P$AT3G51080_01	47231	500	292	2.20E-16
P$AT3G60580_01	28708	464	299	2.20E-16
P$AT3G63350_01	46089	394	268	2.20E-16
P$AT4G36620_01	20563	551	274	2.20E-16
P$AT5G07310_01	84076	363	267	2.20E-16
P$AT5G54070_01	99976	439	284	2.20E-16
P$ATERF14_01	15856	384	262	2.20E-16
P$ATH1_01	13173	536	289	2.20E-16
P$ATHB6_01	130652	570	267	2.20E-16
P$ATHSFA1D_01	46346	554	278	2.20E-16
P$ATMYB15_Q2	62184	565	271	2.20E-16
P$AZF3_01	11800	554	262	2.20E-16
P$BD1_01	15824	388	266	2.20E-16
P$BHLH112_01	3852	492	298	2.20E-16
P$BIM1_02	60296	434	286	2.20E-16
P$BPC1_Q2	269377	498	294	2.20E-16
P$BZR1_01	32722	428	286	2.20E-16
P$C1_Q2	101652	492	289	2.20E-16
P$CBF1_02	6152	412	269	2.20E-16
P$CBF2_03	98817	376	266	2.20E-16
P$CBNAC_01	221922	491	296	2.20E-16
P$CCA1_01	2536	547	280	2.20E-16
P$CDC5_01	2518	472	274	2.20E-16
P$CDF2_01	110198	520	287	2.20E-16
P$CRF1_01	13949	384	260	2.20E-16
P$CRF1_02	13819	383	260	2.20E-16
P$CRF2_01	113608	356	263	2.20E-16
P$CRF3_01	32437	389	268	2.20E-16
P$CRF4_01	86376	352	263	2.20E-16

P$DEAR3_02	166196	380	267	2.20E-16
P$DOF_Q2	80159	509	289	2.20E-16
P$DOF1_01	155359	533	285	2.20E-16
P$DOF2_01	331902	525	289	2.20E-16
P$DOF3_01	331902	525	289	2.20E-16
P$DOF18_01	332107	525	289	2.20E-16
P$DOF24_01	332712	527	289	2.20E-16
P$DOF53_01	332690	527	289	2.20E-16
P$DOF56_01	379892	523	289	2.20E-16
P$DOF57_01	214045	518	291	2.20E-16
P$DREB1A_01	13648	383	260	2.20E-16
P$DREB1A_04	39525	449	274	2.20E-16
P$DREB1B_01	362183	399	273	2.20E-16
P$ERF1_04	13135	439	276	2.20E-16
P$ERF1_05	13819	383	260	2.20E-16
P$ERF1_Q2	88937	360	265	2.20E-16
P$ERF1A_01	12039	392	272	2.20E-16
P$ERF1B_06	99332	358	265	2.20E-16
P$ERF2_01	86386	352	263	2.20E-16
P$ERF2_02	12039	392	272	2.20E-16
P$ERF2_03	69848	351	264	2.20E-16
P$ERF3_04	125601	370	267	2.20E-16
P$ERF4_02	102734	359	264	2.20E-16
P$ERF4_03	13819	383	260	2.20E-16
P$ERF4_04	129023	370	266	2.20E-16
P$ERF5_01	14655	384	262	2.20E-16
P$ERF5_02	6333	449	277	2.20E-16
P$ERF6_01	15912	382	262	2.20E-16
P$ERF7_02	256417	394	271	2.20E-16
P$ERF8_01	250004	393	271	2.20E-16
P$ERF11_01	417066	406	274	2.20E-16
P$ERF13_01	5462	381	261	2.20E-16
P$ERF13_02	233205	391	271	2.20E-16
P$ERF15_01	14057	390	273	2.20E-16
P$ERF039_01	97714	435	276	2.20E-16
P$ERF069_01	507792	409	275	2.20E-16
P$ERF094_01	121944	366	266	2.20E-16
P$ERF096_01	242150	394	271	2.20E-16
P$ERF098_01	242294	394	272	2.20E-16
P$ERF104_01	15824	388	266	2.20E-16
P$ERF105_01	14957	387	261	2.20E-16
P$ERF112_02	189555	382	268	2.20E-16
P$FUS3_01	17848	515	287	2.20E-16
P$GATA1_01	256202	523	278	2.20E-16
P$GATA8_01	272639	478	288	2.20E-16
P$GATA15_01	275446	479	288	2.20E-16

P$GT1_Q6	81801	550	277	2.20E-16
P$GT1_Q6_01	25066	541	269	2.20E-16
P$GT1_Q6_02	28500	561	270	2.20E-16
P$HAT1_01	27502	574	266	2.20E-16
P$HMGIY_01	273517	523	276	2.20E-16
P$HSF3_01	44898	434	279	2.20E-16
P$HSFA1E_01	332681	402	274	2.20E-16
P$HSFA2_01	108474	520	286	2.20E-16
P$HSFA4A_01	53043	561	273	2.20E-16
P$HSFC1_01	84912	528	291	2.20E-16
P$JERF1_01	127518	378	269	2.20E-16
P$KAN1_01	19949	565	273	2.20E-16
P$KNOX3_01	185537	521	284	2.20E-16
P$LEC2_01	43246	531	282	2.20E-16
P$MYB1L_01	45321	496	290	2.20E-16
P$MYBAS1_01	337222	525	284	2.20E-16
P$NAC92_01	57087	505	289	2.20E-16
P$OPBP1_01	22057	388	268	2.20E-16
P$ORA59_01	14057	390	273	2.20E-16
P$OS05G0497200_01	105512	362	264	2.20E-16
P$P_01	17043	461	290	2.20E-16
P$PBF_01	158731	528	286	2.20E-16
P$PBF_Q2	201203	526	288	2.20E-16
P$PBF_Q2_01	332930	527	289	2.20E-16
P$PEND_01	5455	583	264	2.20E-16
P$PHYPA38837_09	332107	525	289	2.20E-16
P$PHYPA140773_01	332107	525	289	2.20E-16
P$PHYPA153324_03	332291	526	289	2.20E-16
P$PIF5_01	6904	480	291	2.20E-16
P$PTI5_01	19441	396	273	2.20E-16
P$PTI6_01	2779	406	276	2.20E-16
P$RAP23_02	87406	353	263	2.20E-16
P$RAP26_03	62868	349	262	2.20E-16
P$RAP26L_01	26342	421	278	2.20E-16
P$RAP26L_02	24121	420	278	2.20E-16
P$RAP210_04	14883	376	264	2.20E-16
P$RAP211_01	13819	383	260	2.20E-16
P$RAV1_01	28135	544	283	2.20E-16
P$RAV1_02	61622	462	285	2.20E-16
P$REM1_01	10871	544	277	2.20E-16
P$RRTF1_01	76182	351	264	2.20E-16
P$RRTF1_02	20445	376	266	2.20E-16
P$RVE1_01	2536	547	280	2.20E-16
P$SED_Q2	121268	517	291	2.20E-16
P$SPL4_01	198031	526	280	2.20E-16
P$TFIIAL_01	56891	555	276	2.20E-16

P$TGA1A_Q2	141301	486	280	2.20E-16
P$TSI1_01	2556	403	276	2.20E-16
P$TSRF1_01	14065	386	271	2.20E-16
P$WRKY8_01	59464	540	279	2.20E-16
P$WRKY15_01	59433	540	279	2.20E-16
P$WRKY18_02	29649	535	280	2.20E-16
P$WRKY21_02	48518	539	279	2.20E-16
P$WRKY40_01	132762	529	283	2.20E-16
P$WRKY43_02	55697	539	278	2.20E-16
P$WRKY48_02	203911	525	284	2.20E-16
P$WRKY57_01	59429	540	279	2.20E-16
P$WRKY75_01	36273	534	281	2.20E-16

Fly, Mouse, Rat, and Human Data Extraction

The *O. sativa* binding sites collected above were inputted into the MATCH program. The MATCH settings were configured to search for similar non-plant eukaryotic binding sites against the TRANSFAC database. Accordingly, an immense collection of non-plant eukaryotic species TFBSs was generated from TRANSFAC using the MATCH program. The non-plant eukaryotic species are comprised of fruit flies, mice, rats, and humans. The resulting file name is *allspecies.csv* and is 7.34 GB in size. As a result of its gigantic size, the file was split into fourteen 524.3 MB sized files in order to extract data without significantly long execution times.

The contents within the file include the names of the motifs, the binding site positions, and the sequences that characterize the motifs. Like the previous data file, the average binding site relative to the nearest TSS was calculated for all motifs by implementing scripts programmed in Perl. The names of the programs developed to do so are all_species_tfbs_extraction.pl and TFBS_average.pl. Furthermore, the consensus sequence motif was computed for the average binding sites positioned near a TSS. Since each motif name is represented by thousands of binding sites, the consensus sequence was calculated by taking the letter that appears most frequently at each position of the sequence. The programs concocted to compute the consensus

sequence motif are *all_species_motif_extraction.pl* and *base_average.pl*. Finally, a sequence logo for each of the consensus sequences was generated using the WebLogo 3 software.

Program 2: all_species_tfbs_extraction.pl

This program calculates the average nucleotide binding site position relative to the closest TSS for a given data file and sequence motif when executed. The required inputs include the name of the file with the binding site data and the name of the motif being searched. The binding site positions relative to the start of each 1000 nucleotide long mRNA fragment (TSS+500, TSS-500) are listed in the file. The nearest TSS is located 500 nucleotides away from the beginning of each 1000 nucleotide long mRNA transcript. Accordingly, 500 is subtracted from the binding site position in order to calculate the distance between the TSS and binding site. Once all binding site positions for each sequence motif have been computed, the average is taken. The program is executed fourteen times for each motif name because of the partitioning of the colossal allspecies.csv file. The binding site positions are written out to a file each time the script is executed.

Program 3: TFBS_average.pl

Due to its enormous size, the allspecies.csv file was split into fourteen smaller files in order to reduce lag time and increase efficiency. The output file for the all_species_tfbs_extraction.pl program contains the binding site positions and is the input required to run this script. The execution of this program results in the averaging of the binding site positions computed from the all_species_tfbs_extraction.pl program. The output is written out to a file.

A list of fly, mouse, rat, and human motifs along with their average positions relative to the nearest TSS are shown in the Table 4. The average distances in table 4 are averages computed between all of the non-plant eukaryotic species. In other words, the binding site

distances for fly, mouse, rat, and human species were added together in order to compute the average distances in table 4. The sequence names and species ID for each binding site are presented in the supplementary materials. Each motif name consists of thousands of TFBSs whose average position relative to the closest TSS is calculated between fly, mouse, rat, and human species. The first column lists the motif names while the second column presents the average binding site position relative to the nearest TSS.

Table 4. Average distances to the nearest TSS for binding sites recognized by various transcription factors of flies, mice, rats, and humans within eukaryotic genomes

Fly, mouse, rat, and human motif name	Number of binding sites	Average binding site distance relative to nearest TSS	Standard deviation for average binding site distance (+/-)	P-value
P$ABI3_01	11651	258	142	2.20E-16
P$AHL20_02	780690	260	139	2.20E-16
P$AHL25_01	983202	260	140	2.20E-16
P$ALFIN1_Q2	34777	239	142	2.20E-16
P$ARF8_01	51595	249	144	2.20E-16
P$ARR1_01	79315	254	143	2.20E-16
P$ARR10_01	38177	255	143	2.20E-16
P$ASR1_01	161254	246	140	2.20E-16
P$AT1G21910_01	622214	229	142	2.20E-16
P$AT1G26590_01	362201	258	141	2.20E-16
P$AT1G26610_01	229434	249	144	2.20E-16
P$AT1G49120_01	1464	211	137	1.69E-03
P$AT1G53910_01	14313	211	143	2.20E-16
P$AT1G67970_01	357417	249	143	2.20E-16
P$AT1G68550_01	7295	204	142	0.0325
P$AT2G15660_01	15326	262	141	2.20E-16
P$AT2G38090_01	386	240	139	2.19E-08
P$AT2G41690_01	139681	231	146	2.20E-16
P$AT2G47520_01	15667	212	143	2.20E-16
P$AT3G18650_01	1817	255	143	2.20E-16
P$AT3G25890_01	2742	211	143	8.63E-05
P$AT3G51080_01	36416	252	144	2.20E-16
P$AT3G60580_01	19578	255	140	2.20E-16

P$AT3G63350_01	20571	217	140	2.20E-16
P$AT4G36620_01	6542	249	143	2.20E-16
P$AT5G07310_01	5870	206	144	6.79E-04
P$AT5G54070_01	180041	233	144	2.20E-16
P$ATERF14_01	6571	203	141	0.0989
P$ATH1_01	11443	241	148	2.20E-16
P$ATHB6_01	137667	262	140	2.20E-16
P$ATHSFA1D_01	59837	255	143	2.20E-16
P$ATMYB15_Q2	18829	252	142	2.20E-16
P$AZF3_01	30089	265	136	2.20E-16
P$BD1_01	945	216	142	4.10E-04
P$BHLH112_01	656	263	140	2.20E-16
P$BIM1_02	88651	246	143	2.20E-16
P$BPC1_Q2	30082	255	143	2.20E-16
P$BZR1_01	9194	236	145	2.20E-16
P$C1_Q2	98129	248	144	2.20E-16
P$CBF1_02	3831	211	142	8.82E-07
P$CBF2_03	7407	198	141	0.139
P$CBNAC_01	19109	244	145	2.20E-16
P$CCA1_01	378110	257	141	2.20E-16
P$CDC5_01	8265	224	144	2.20E-16
P$CDF2_01	50112	252	143	2.20E-16
P$CRF1_01	642554	228	142	2.20E-16
P$CRF1_02	2709	211	137	1.25E-05
P$CRF2_01	15415	205	142	6.47E-05
P$CRF3_01	663823	229	143	2.20E-16
P$CRF4_01	18182	206	142	2.07E-09
P$DEAR3_02	1070132	233	143	2.20E-16
P$DOF_Q2	31679	252	143	2.20E-16
P$DOF1_01	23109	252	143	2.20E-16
P$DOF2_01	22626	250	143	2.20E-16
P$DOF3_01	28393	249	143	2.20E-16
P$DOF18_01	487762	253	142	2.20E-16
P$DOF24_01	400831	255	142	2.20E-16
P$DOF53_01	606992	254	142	2.20E-16
P$DOF56_01	223083	255	142	2.20E-16
P$DOF57_01	1032498	253	142	2.20E-16
P$DREB1A_01	1156	216	136	6.51E-05
P$DREB1A_04	4127	227	146	2.20E-16
P$DREB1B_01	49008	219	143	2.20E-16
P$ERF1_04	2150	225	147	1.09E-14
P$ERF1_05	4442	198	141	0.282
P$ERF1_Q2	11263	212	143	2.20E-16
P$ERF1A_01	821002	230	143	2.20E-16
P$ERF1B_06	14836	211	144	2.20E-16
P$ERF2_01	50032	215	143	2.20E-16

P$ERF2_02	820652	230	143	2.20E-16
P$ERF2_03	10507	211	143	1.33E-14
P$ERF3_04	19180	212	143	2.20E-16
P$ERF4_02	10810	207	142	2.84E-07
P$ERF4_03	2078	211	138	3.78E-04
P$ERF4_04	51039	213	142	2.20E-16
P$ERF5_01	844628	231	143	2.20E-16
P$ERF5_02	1562	231	145	2.20E-16
P$ERF6_01	767612	230	143	2.20E-16
P$ERF7_02	326913	226	144	2.20E-16
P$ERF8_01	289921	228	144	2.20E-16
P$ERF11_01	390390	227	144	2.20E-16
P$ERF13_01	599375	228	142	2.20E-16
P$ERF13_02	16396	213	143	2.20E-16
P$ERF15_01	586403	230	143	2.20E-16
P$ERF039_01	52188	219	142	2.20E-16
P$ERF069_01	307192	225	144	2.20E-16
P$ERF094_01	9359	207	143	5.20E-06
P$ERF096_01	109142	220	143	2.20E-16
P$ERF098_01	470700	228	144	2.20E-16
P$ERF104_01	661531	230	142	2.20E-16
P$ERF105_01	7311	203	142	0.0411
P$ERF112_02	90061	221	142	2.20E-16
P$FUS3_01	2048	239	140	2.20E-16
P$GATA1_01	83567	261	139	2.20E-16
P$GATA8_01	73609	254	143	2.20E-16
P$GATA15_01	82319	252	144	2.20E-16
P$GT1_Q6	52925	255	143	2.20E-16
P$GT1_Q6_01	40148	266	136	2.20E-16
P$GT1_Q6_02	30507	255	142	2.20E-16
P$HAT1_01	2040	264	137	2.20E-16
P$HMGIY_01	115212	259	140	2.20E-16
P$HSF3_01	32615	237	138	2.20E-16
P$HSFA1E_01	286267	231	143	2.20E-16
P$HSFA2_01	172992	249	143	2.20E-16
P$HSFA4A_01	87073	254	144	2.20E-16
P$HSFC1_01	311185	250	144	2.20E-16
P$JERF1_01	16148	212	143	2.20E-16
P$KAN1_01	1085519	251	142	2.20E-16
P$KNOX3_01	36235	237	145	2.20E-16
P$LEC2_01	2835	255	140	2.20E-16
P$MYB1L_01	52601	253	141	2.20E-16
P$MYBAS1_01	34130	248	144	2.20E-16
P$NAC92_01	38268	248	143	2.20E-16
P$OPBP1_01	1912	209	146	6.99E-03
P$ORA59_01	674679	231	143	2.20E-16

P$OS05G0497200_01	2491	190	142	5.41E-04
P$P_01	23564	243	142	2.20E-16
P$PBF_01	22793	252	143	2.20E-16
P$PBF_Q2	28940	254	143	2.20E-16
P$PBF_Q2_01	36914	254	143	2.20E-16
P$PEND_01	6278	264	140	2.20E-16
P$PHYPA38837_09	351865	253	142	2.20E-16
P$PHYPA140773_01	364898	254	142	2.20E-16
P$PHYPA153324_03	249086	253	143	2.20E-16
P$PIF5_01	1026	239	149	2.20E-16
P$PTI5_01	8020	204	142	6.98E-03
P$PTI6_01	457	217	144	0.0119
P$RAP23_02	1225043	233	143	2.20E-16
P$RAP26_03	587236	225	143	2.20E-16
P$RAP26L_01	714209	233	143	2.20E-16
P$RAP26L_02	1318	215	147	1.43E-04
P$RAP210_04	690187	228	143	2.20E-16
P$RAP211_01	596911	229	142	2.20E-16
P$RAV1_01	30569	248	144	2.20E-16
P$RAV1_02	48840	250	141	2.20E-16
P$REM1_01	573229	251	143	2.20E-16
P$RRTF1_01	7633	205	143	2.81E-03
P$RRTF1_02	728079	227	143	2.20E-16
P$RVE1_01	400512	257	141	2.20E-16
P$SED_Q2	13626	248	143	2.20E-16
P$SPL4_01	155406	249	143	2.20E-16
P$TFIIAL_01	765733	254	142	2.20E-16
P$TGA1A_Q2	11028	232	145	2.20E-16
P$TSI1_01	422	216	143	0.0194
P$TSRF1_01	6737	202	142	0.159
P$WRKY8_01	51231	245	145	2.20E-16
P$WRKY15_01	43313	243	145	2.20E-16
P$WRKY18_02	3074	254	142	2.20E-16
P$WRKY21_02	47722	245	145	2.20E-16
P$WRKY40_01	81397	244	145	2.20E-16
P$WRKY43_02	103118	250	144	2.20E-16
P$WRKY48_02	111730	243	144	2.20E-16
P$WRKY57_01	57382	249	144	2.20E-16
P$WRKY75_01	1969	253	143	2.20E-16

Table 5 lists the fly, mouse, rat, and human TF binding sites that are located in close proximity to the nearest TSS of the gene they regulate.

Table 5. Transcription factors of flies, mice, rats, and humans, which, on average, bind DNA in close proximity (within 200 nts) to the nearest TSS.

Fly, mouse, rat, and human motif name	Number of binding sites	Average binding site distance relative to nearest TSS	Standard deviation for average binding site distance (+/-)	P-value
P$AT1G68550_01	7295	204	142	0.0325
P$ATERF14_01	6571	203	141	0.0989
P$CBF2_03	7407	198	141	0.139
P$ERF1_05	4442	198	141	0.282
P$ERF105_01	7311	203	142	0.0411
P$PTI6_01	457	217	144	0.0119
P$TSI1_01	422	216	143	0.0194
P$TSRF1_01	6737	202	142	0.159

Program 4: all_species_motif_extraction.pl

The execution of this program results in the computation of the nucleotides that comprise a transcription factor binding site. This is accomplished by examining a binding site and marking the nucleotide that appears at each position. When all binding sites for a particular motif have been examined, the consensus nucleotide at each position is ready to be computed. The required inputs are the file name with the binding site data and the name of the motif being examined. The output is written to a file and contains the nucleotide at each position of all binding sites for the motif being analyzed.

Program 5: base_average.pl

Execution of this program results in the calculation of the consensus sequence motif. The input for this program is the output for the *all_species_motif_extraction.pl* code. The nucleotides that occur most frequently at each position will be incorporated into the consensus motif

sequence. The output is printed to a file and contains the consensus nucleotides at each position along with the final consensus sequence motif.

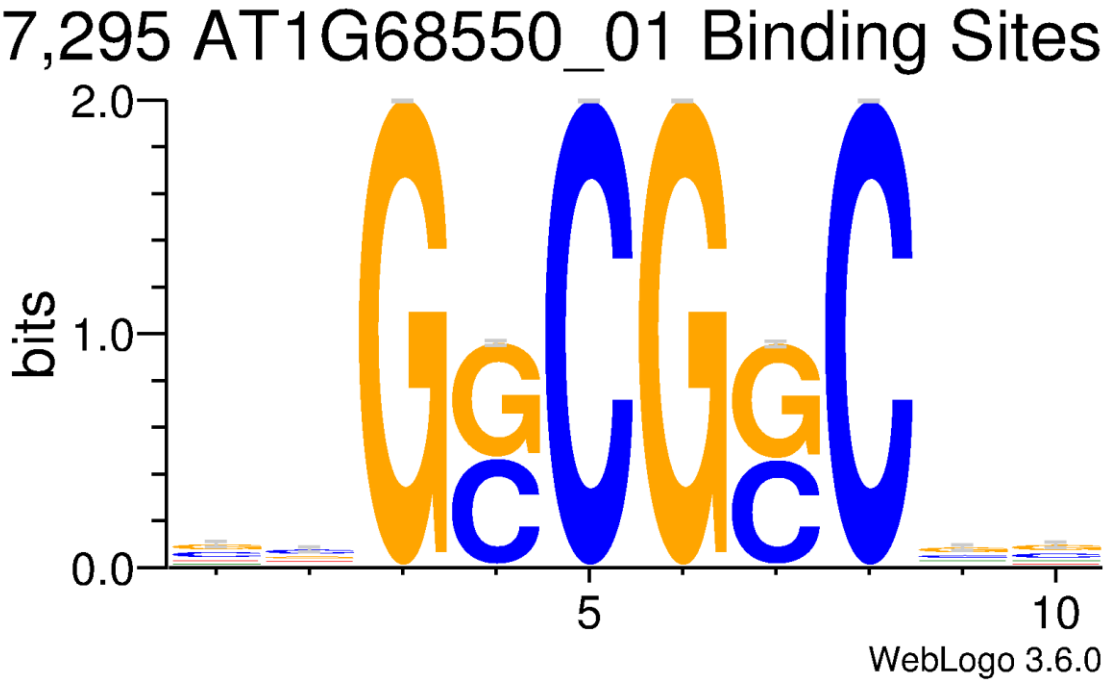
The consensus sequence motifs for the rice, fly, mouse, rat, and human binding sites located near the closest TSS were computed and are displayed in the Table 6. Each motif name consists of thousands of TFBSs whose average nucleotide at each position of its sequence is calculated. The first column lists the motif name while the second column presents the actual consensus sequence that defines a binding site.

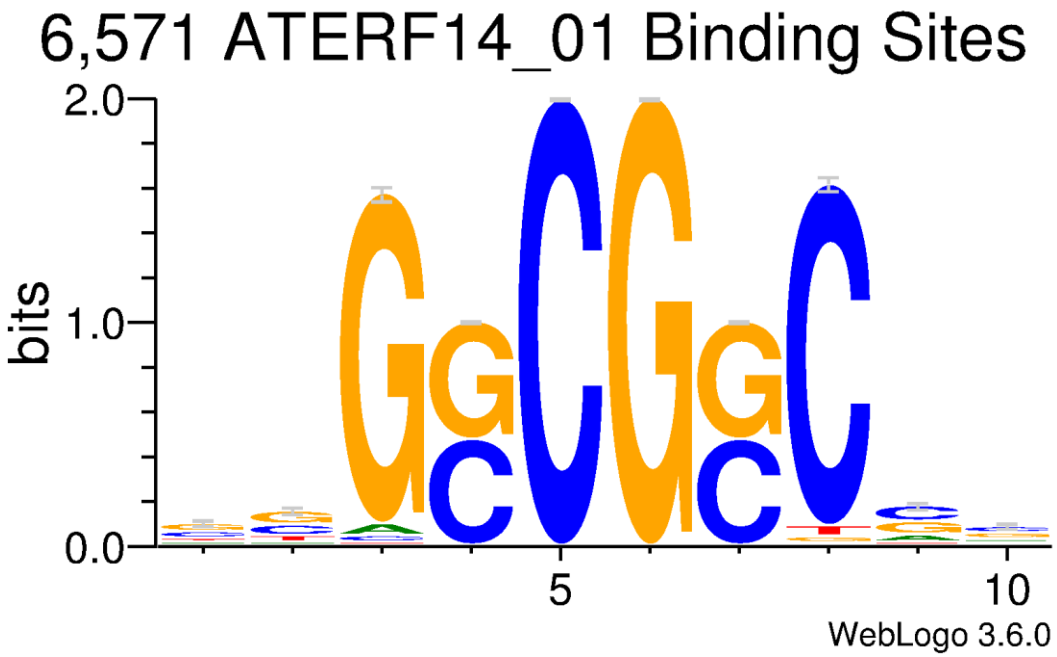
Table 6. Consensus sequence motifs for *O. sativa*, fly, mouse, rat, and human binding sites located, on average, in close proximity (within 200 nts) to the nearest TSS.

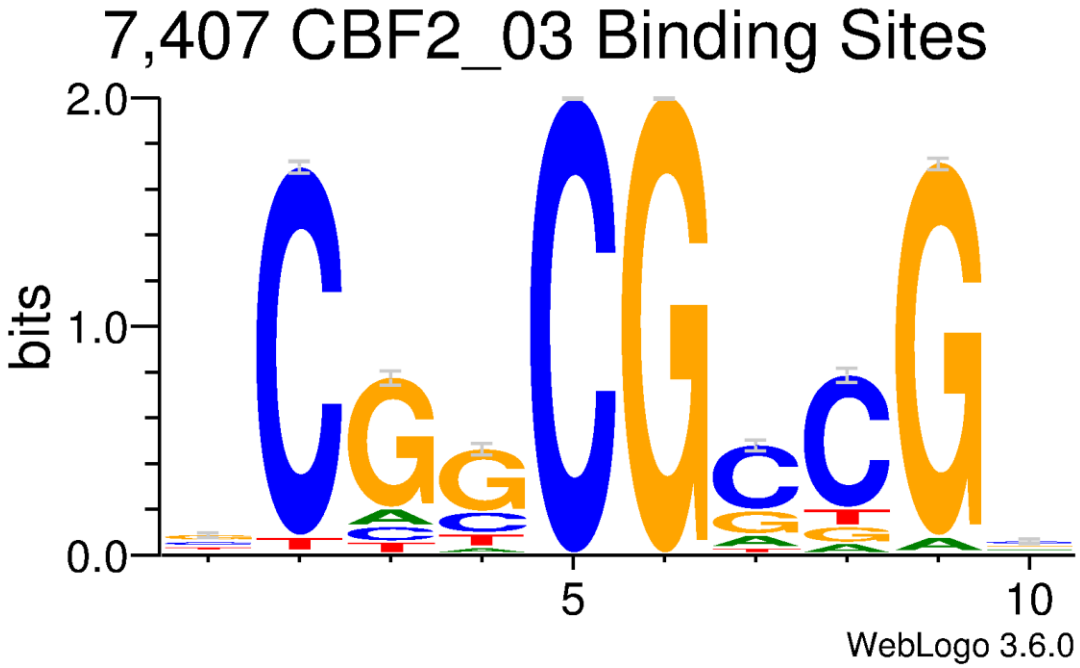
Motif name	Motif consensus sequence
P$AT1G68550_01	ccgGCGGCgg
P$ATERF14_01	gggGCGGCcc
P$CBF2_03	gCgGCGCcGc
P$ERF1_05	gcgGCGGCgg
P$ERF105_01	ccgGCGGCgg
P$PTI6_01	gtggCGGCCg
P$TSI1_01	ctggCGGCCg
P$TSRF1_01	ccgGCGGCgg

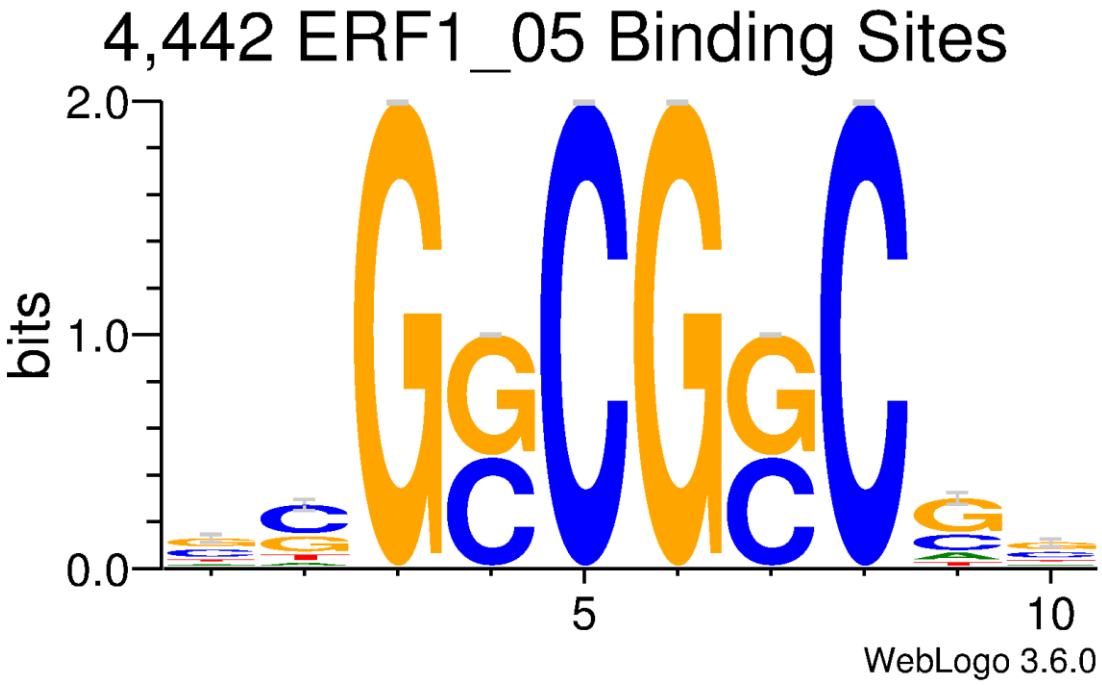
The sequence logos for the Rice *O. sativa*, Fly, Mouse, Rat, and Human P$AT1G68550_01, P$ATERF14_01, P$CBF2_03, P$ERF1_05, P$ERF105_01, P$PTI6_01, P$TSI1_01, and P$TSRF1_01 binding site motifs are shown below in Figure 3. The x-axis displays the nucleotide position for the motif sequence while the y-axis represents the information content in bits, calculated from the formula, $R_{seq} = S_{max} - S_{obs} = \log_2 N - (-\Sigma_{n=1}\ p_n \log_2 p_n)$. P_n represents the observed frequency of symbol n for a specific position while N is the amount of unique characters for a given sequence type (Crooks, et al., 2004). Additionally, the size of each

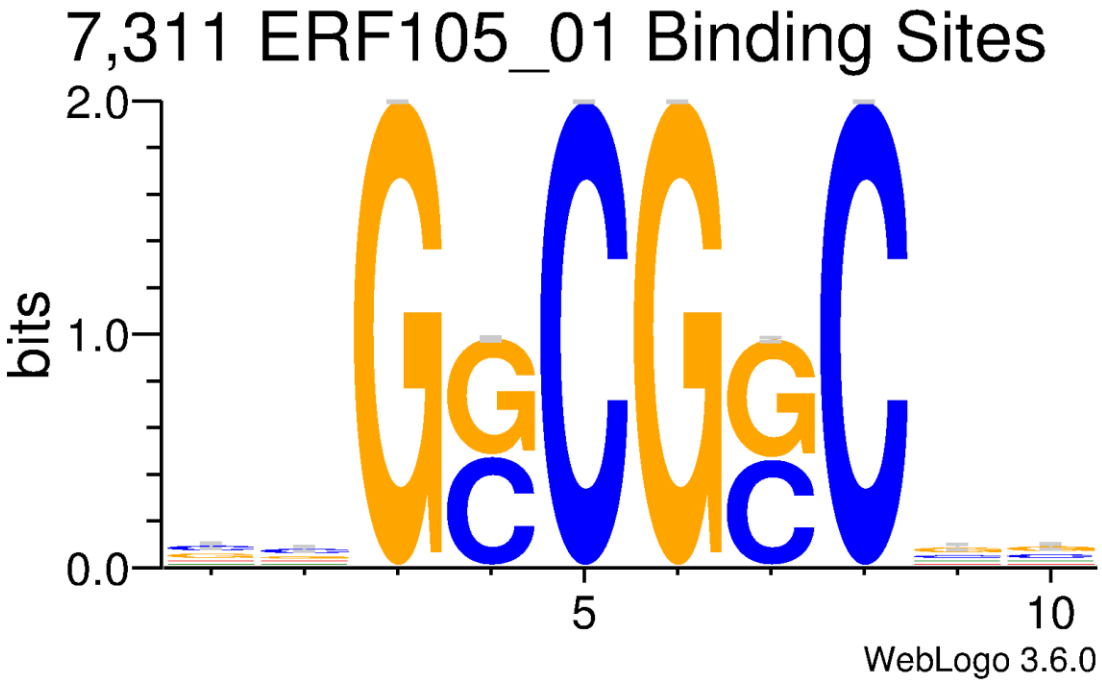
individual letter represents its frequency at that particular position while the overall height of each stack personifies the information content.

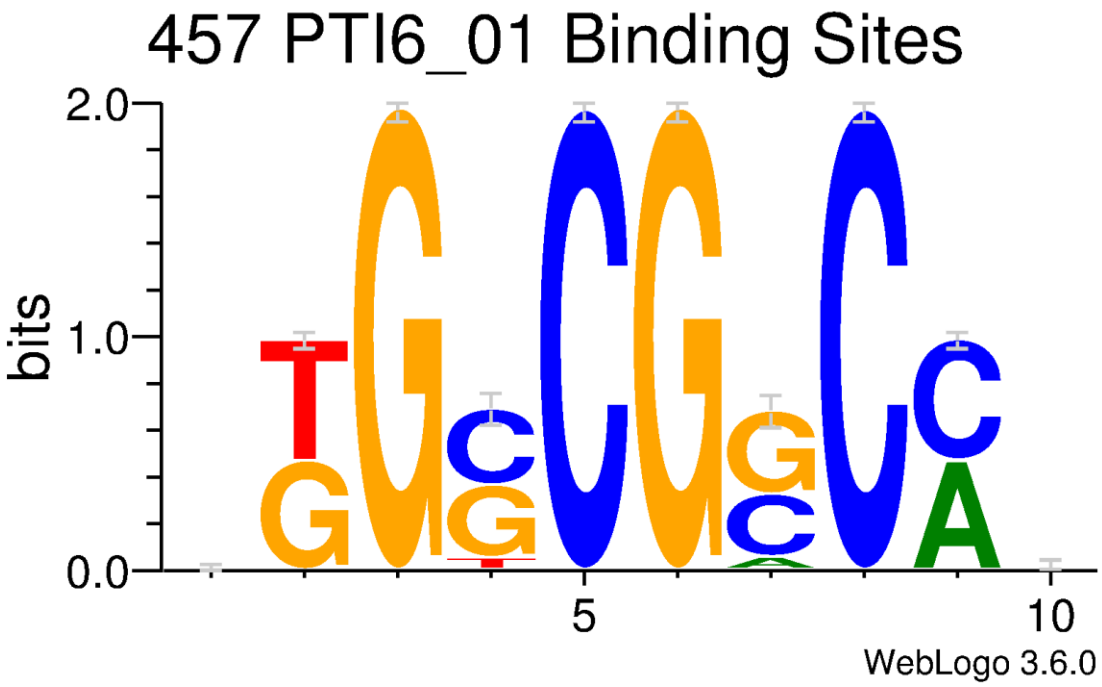


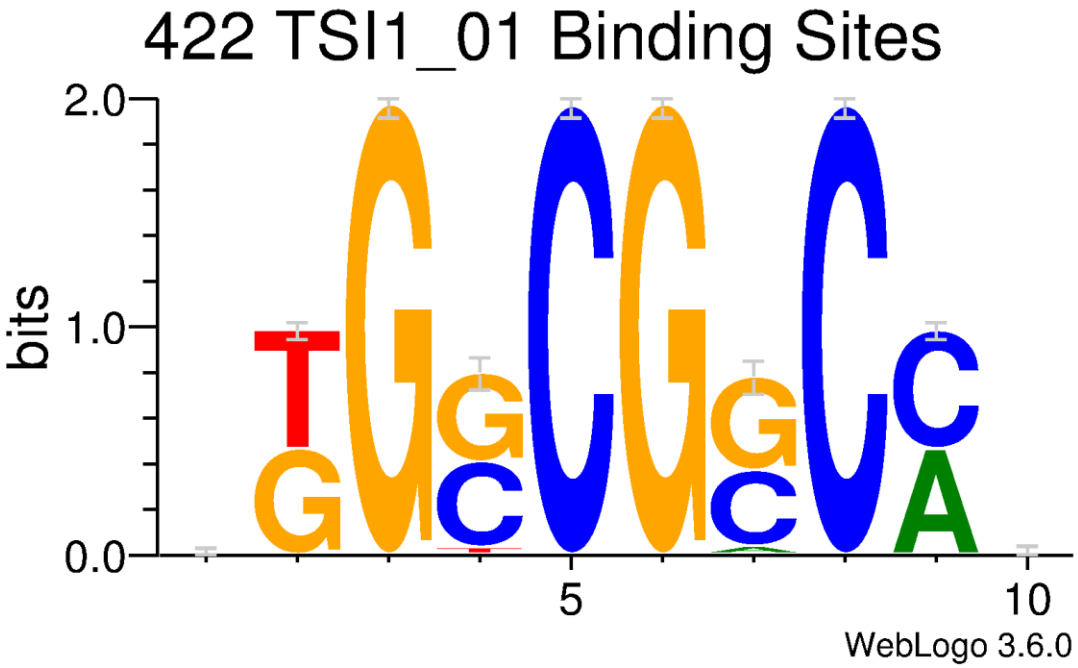

6,571 ATERF14_01 Binding Sites
bits
2.0
1.0
0.0
5
10
WebLogo 3.6.0


7,407 CBF2_03 Binding Sites
bits
2.0
1.0
0.0
5
10
WebLogo 3.6.0


4,442 ERF1_05 Binding Sites
bits
2.0
1.0
0.0
5
10
WebLogo 3.6.0


7,311 ERF105_01 Binding Sites
bits
2.0
1.0
0.0
5
10
WebLogo 3.6.0


457 PTI6_01 Binding Sites
2.0
1.0
0.0
bits
5
10
WebLogo 3.6.0


422 TSI1_01 Binding Sites
2.0
1.0
0.0
bits
5
10
WebLogo 3.6.0

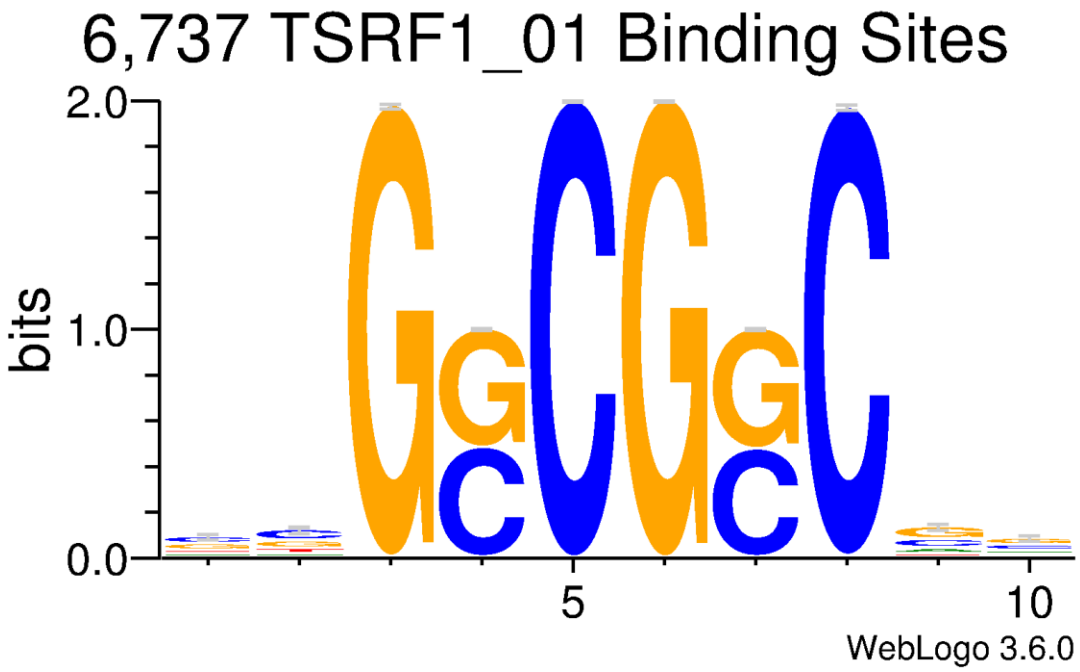


Figure 3. Sequence logos for rice *O. sativa*, fly, mouse, rat, and human binding sites located in close proximity to the nearest TSS.

DISCUSSION

O. sativa Orthologous Transcription Factors

As can be seen in Table 1, a total of 40 non-plant eukaryotic transcription factors orthologous to *O. sativa* TFs were discovered in several species including humans, mice, chickens, and fruit flies. Half of the orthologous transcription factors were found in human genome, while the other half come from rodents and insects. A majority of the orthologous transcription factor templates had their three-dimensional structures resolved through x-ray crystallography, while the rest were defined using Nuclear Magnetic Resonance Spectroscopy (NMR) or cryogenic electron microscopy (cryo-EM) (Table 2). The sequence identities and sequence similarities used to define homology for the transcription factors range from 17-73% and 0.275-0.530, respectively. The coverage is another indicator of homology and is shown to have a range of 0.105-0.556 (Table 2).

In addition to the pairwise sequence alignments, the *O. sativa* transcription factors and their orthologues were also evaluated for similarities in their structural characteristics and biological roles. As example, the *O. sativa* MADS3 gene, which codes for the MADS-box transcription factor 3 protein, was compared to its human and mouse orthologs, the Myocyte-specific enhancer factor 2B (MEF2C gene) and myocyte-specific enhancer factor 2C (MEF2B gene). A protein structure alignment of the MADS3 and MEF2 proteins was conducted using the jFATCAT flexible program. According to data depicted at Figure 2, there is an overall 87.14% structure identity and 95.71 structure similarity. According to Uniprot, the MADS-box

transcription factor 3 protein is involved in the development of floral organs in *O. sativa*, while the myocyte enhancement factor 2 transcription factor in humans and mice is essential for muscle cell differentiation. Hence, both transcription factors play a role of specific tissue development.

O. sativa Transcription Factor Binding Sites

A total of 154 *O. sativa* transcription factor motifs were obtained from the TRANSFAC database and evaluated for their average binding site positions relative to the nearest TSS. The entire list of *O. sativa* motif names and their average binding site positions relative to the closest TSS are presented in Table 3. According to Table 3, none of the 154 collection of *O. sativa* binding sites appear to be located in close proximity, which is defined as within 200 nt, to the nearest TSS. This is because the average distances from the *O. sativa* binding sites to the nearest TSS range from 349 nts to 583 nts for the motifs shown in Table 3. Additionally, the p-values calculated for each collection of *O. sativa* TFBSs from a one-sample t-test are approximately 0, which means there is a high degree of confidence that the average binding site distances are not within 200 nts of the nearest TSS. TFBSs located in close proximity to known TSSs may be used for prediction of TSSs in the less characterized genes.

Although there weren't any *O. sativa* TFBSs within 200 nts of the nearest TSS, each binding site still contributes significantly to the functioning of *O. sativa* cells. In plants, the ABI3 TF is involved in regulating seed development along with leaf and embryo degreening (Léon-Kloosterziel, et al., 1996; Okamoto, et al., 2010). The AT3G60580 TF is also known as Zinc finger protein 9 (ZAT9) and may play a role in various responses to stress (Feurtado, et al., 2011; Giri, et al., 2011). The aborted microspores (AMS) TF is vital for tapetum development, male fertility, and pollen differentiation (Sorenson, et al., 2003). The BES1-interacting Myc-like protein 1 (BIM1) is a positive brassinosteroid-signaling TF that plays a role in various growth and

developmental processes like cell elongation, vascular development, and stress responses (Yin, et al., 2005).

Fly, Mouse, Rat, and Human Transcription Factor Binding Sites

The fly, mouse, rat, and human equivalents of the 154 *O. sativa* transcription factor motifs, and the collection of their binding sites were extracted from the TRANSFAC database. For each TF binding site, these datasets were used to calculate average distance from that site to the closest TSS. The complete list of motifs with their average distances relative to the nearest TSS and standard deviations is displayed in Table 4. In addition, a p-value was computed for each set of average distances by conducting a one-sample t-test with a 95% confidence interval. On average, the binding sites were located within a range of 190 nts to 265 nts from the closest TSS for the fly, mouse, rat, and human motifs.

P$AT1G68550_01, P$ATERF14_01, P$CBF2_03, P$ERF1_05, P$ERF105_01, P$PTI6_01, P$TSI1_01, and P$TSRF1_01 matrices all have relatively short average distances from the binding site to the nearest TSS. Peculiarly, TFs corresponding to matrices P$AT1G68550_01, P$ATERF14_01, P$CBF2_03, P$ERF1_05, P$ERF105_01, P$PTI6_01, P$TSI1_01, and P$TSRF1_01 all belong to the ERF family of TFs (Ohme-Takagi and Shinshi, 1995; Suzuki, et al., 1998), which is involved in bolstering the defense of plants by activating the stress signaling pathways. TF proteins of ERF family attach to the GCC-box pathogenesis-related promoter element (Eyal, et al., 1993), and, by that, modulate the genes comprising the stress signal transduction pathways (Fujimoto et al., 2000). The average distance for the P$AT1G68550_01 TFBS was 204±142 nts, with a p-value of 0.0325.

P$AT1G68550_01 is a TFBSsfor the Ethylene-responsive transcription factor ERF118, which functions as a transcriptional activator. P$ATERF14_01 (ERF14), P$ERF1_05, and

P$ERF105_01 are also collections of binding sites for ethylene-responsive transcription factors that function as activators of transcription.

P$ERF105_01 has an average distance of 203±142 nts and a p-value of 0.0411, which provides a moderate degree of confidence that the mean is located within 200 nts of the nearest TSS. ERF14 offers pathogen resistance (Sanchez, et al., 2007), ERF1 augments the response to various stresses like salt, drought, and heat (Yang, et al., 2009), and the ERF105 TF provides resistance to cold temperatures (Bolt, et al., 2017).

P$CBF2_03 are the TFBSs for the dehydration-responsive element-binding protein 1C (DREB1C) that binds specifically to the DNA sequence 5'-[AG]CCGAC-3'. CBF/DREB1 TFs are activators that play a major role in freezing tolerance and cold acclimation by binding to the C-repeat/DRE element and facilitating cold-inducible transcription (UniProt) (Agarwal, et al., 2006). P$TSI1_01 has an average distance of 216±143 nts and a p-value of 0.0194, which provides a small degree of confidence that the mean is located within 200 nts of the nearest TSS. P$TSI1_01 is the binding site for the tobacco stress-induced 1 TF that may play a role in curtailing abiotic and biotic stress by stimulating signal transduction pathways (Park, et al., 2001). The average distance for P$PTI6_01 is 217±144 nts along with a p-value of 0.0119, which provides a small degree of confidence that the mean is positioned within 200 nts of the nearest TSS. P$PTI6_01 is the collection of TFBSs for the pathogenesis-related genes transcriptional activator PTI6 which stimulates the protective genes of plants in order to increase resistance against microbial and fungal pathogens (Gu, et al., 2002).

Comparative Analysis of Consensus Sequence Motifs

The consensus sequence motifs for the 8 TFs of *O. sativa*, fly, mouse, rat, and human with an average location in close proximity of their binding sites to the nearest TSS were

calculated for comparative analysis and can be observed in table 6. The P$AT1G68550_01, P$ATERF14_01, P$CBF2_03, P$ERF1_05, P$ERF105_01, P$PTI6_01, P$TSI1_01, and P$TSRF1_01 binding sites, which are all located very close to the nearest TSS, have consensus sequences of ccgGCGGCgg, gggGCGGCcc, gCgGCGCcGc, gcgGCGGCgg, ccgGCGGCgg, gtggCGGCCg, ctggCGGCCg, and ccgGCGGCgg, respectively.

The corresponding sequence logos for the P$AT1G68550_01, P$ATERF14_01, P$CBF2_03, P$ERF1_05, P$ERF105_01, P$PTI6_01, P$TSI1_01, and P$TSRF1_01 binding sites are displayed in figure 3 and provide additional information in regards to the nucleotide conservation at each position. For instance, the nucleotide letters at positions 3, 5, 6, and 8 in the P$AT1G68550_01 sequence logo are more pronounced than the other positions because they appear more often and possess more information content as a result. Consequently, these particular positions can be considered more evolutionary conserved which typically indicates a significant biological function. Positions 4 and 7 are each split between two letters, which appear to be competing for the positions as a result of selection pressures from the evolutionary process. Positions 3, 5, 6, and 8 for the P$ATERF14_01 binding site carry far more information content than the other positions. In fact, the letters at each of these positions carry a high degree of conservation and closely resemble those of the P$AT1G68550_01 sequence logo. In contrast, positions 4 and 7 carry half the amount of information and are split between G and C nucleotides. The sequence logo for the P$CBF2_03 binding site displays a high degree of conservation and information content at positions 2, 5, 6, and 9 while the remaining positions carry a smaller degree of conservation and information. Positions 5 and 6 each possess two letters that are similar in size, which indicates a similarity in frequency and sequence conservation. Position 5 is split between C and A while position 6 is split between T and G. As a result, both positions are under strong selection pressures for either nucleotide. The sequence logos for P$ERF1_05,

P$ERF105_01, and P$TSRF1_01 mirror that of P$AT1G68550_01 which can be observed from positions 3, 5, 6, and 8. Lastly, the sequence logos for the P$PTI6_01 and P$TSI1_01 binding sites mirror each other at all positions. In conclusion, the consensus and sequence logos for each of the 8 TFBSs analyzed above display a predilection for the GCC-box pathogenesis-related promoter motif, which is a characteristic feature of the ERF family of TFs.

CONCLUSION

The first objective for this study was to determine if promoter-specific and 5'UTR-specific TFs of *Oryza sativa* have orthologous TFs in non-plant eukaryotic species. This aim was successfully accomplished, as a plethora of non-plant eukaryotic orthologous TFs were found through the use of several online tools and databases including UniProt, SWISS-MODEL, and PDB. The orthologues are encoded by genomes of a diverse set of species such as *Drosophila melanogaster* (fruit fly), *Gallus gallus* (chicken), *Mus musculus* (mouse), and *Homo sapien* (human). Among all orthologous TFs discovered, a majority of the orthologs were detected in humans, possibly due to the fact that human genome is relatively well annotated. Thus, it can be concluded that *Oryza sativa* TFs possess orthologues in non-plant eukaryotic species that most likely descended from a common ancestor and diverged after an ancient speciation event.

After successfully locating non-plant eukaryotic TFs orthologous to *Oryza sativa*, the next aim was to compare the TF binding sites by analyzing the distribution of their positions relative to the nearest TSSs in respective genomes. For rice, there were no collections of binding sites with an average location significantly close to or within 200 nts of the nearest TSS. For the non-plant eukaryotes, the TFs with the collection of binding sites located significantly close to or within 200 nts of a TSS are P$AT1G68550_01, P$ATERF14_01, P$CBF2_03, P$ERF1_05, P$ERF105_01, P$PTI6_01, P$TSI1_01, and P$TSRF1_01. In addition, the analysis of TF binding site distributions revealed *Oryza sativa* average binding site locations ranging from 349 nts to 583 nts of the nearest TSS. On the other hand, the binding sites belonging to fruit flies, mice, rats, and humans were found at the distances ranging from 190 nts to 265 nts of the nearest

TSS. Therefore, a positional preference in TF binding may also exist in non-plant eukaryotic species. Further studies on TFBS distributions across eukaryotic species are warranted in order to elucidate the evolutionary mechanisms involved in gene regulation. Because the averages for the binding sites were between the non-plant eukaryotic species, further studies need to be conducted in order to compute average distances relative to the nearest TSS for individual non-plant eukaryotic species. Analysis of the average binding sites needs to be computed Moreover, additional research needs to be conducted in order to determine the exact function for each site and how transcription is activated or deactivated as a result of TFs binding to the sites.

The final goal for this research was to compare the sequence motifs of the binding sites located in close proximity to the nearest TSS, analyze the distributions of nucleotides at each position and detect potential base differences across species. Across 8 selected TFs with average binding sites located significantly close to or within 200 nts of the nearest TSS, sequence motifs were compared between species. The consensus sequences and sequence logos for these TF binding sites provided information content, in bits, for each position of the sequences and revealed a significant degree of nucleotide conservation at particular positions. This was exemplified through the comparison of sequence logos for the P$AT1G68550_01, P$ATERF14_01, P$CBF2_03, P$ERF1_05, P$ERF105_01, P$PTI6_01, P$TSI1_01, and P$TSRF1_01 binding sites, which revealed a high degree of nucleotide conservation at specific positions of the sequence motifs. Thus, it can be concluded that these TF binding sites have been conserved over time and have evolved independently across several eukaryotic species. Such a high degree of conservation also leads to the conclusion that these TFBSs and their distributions are pertinent to specific biological function these factors perform in rice, fruit flies, mice, rats, and humans.

REFERENCES

1. Triska M, Solovyev V, Baranova A, Kel A, Tatarinova TV. Nucleotide patterns aiding in prediction of eukaryotic promoters. *PLoS One*. 2017;12(11):e0187243.

2. Halazonetis TD, Kandil AN. Determination of the c-MYC DNA-binding site. *Proc Natl Acad Sci U S A*. 1991;88(14):6162–6166.

3. Alberts B, Johnson A, Lewis J, et al. How Genetic Switches Work. Molecular Biology of the Cell. 4th edition. *New York: Garland Science*; 2002.

4. Cooper GM. The Complexity of Eukaryotic Genomes. The Cell: A Molecular Approach. 2nd edition. *Sunderland (MA): Sinauer Associates*; 2000.

5. Eckardt NA. Unexpected Structure of Plant Promoters. *Plant Cell*. 2014;26(7):2726.

6. Stepanova M, Tiazhelova T, Skoblov M, Baranova A. A comparative analysis of relative occurrence of transcription factor binding sites in vertebrate genomes and gene promoter areas. *Bioinformatics*. 2005;21(9):1789–1796.

7. Schneider TD. Consensus sequence Zen. *Appl Bioinformatics*. 2002;1(3):111–119.

8. Shahmuradov IA, Umarov RK, Solovyev VV. TSSPlant: a new tool for prediction of plant Pol II promoters. *Nucleic Acids Res*. 2017;45(8):e65.

9. Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume D. A. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Reviews Genetics*. 2007;8:424.

10. Troukhan M, Tatarinova T, Bouck J, Flavell RB, Alexandrov NN. Genome-wide discovery of cis-elements in promoter sequences using gene expression. *OMICS: A Journal of Integrative Biology*. Apr 2009.

11. Tatarinova TV, Alexandrov NN, Bouck JB, Feldmann KA. GC_3 biology in corn, rice, sorghum and other grasses. *BMC Genomics*. 2010; 11:308.

12. Yella VR, Bansal M. DNA structural features of eukaryotic TATA-containing and TATA-less promoters. *FEBS Open Bio*. 2017; 7(3), 324-334.

13. Singh S, Kaur S, Goel N. A Review of Computational Intelligence Methods for Eukaryotic Promoter Prediction. *Nucleosides, Nucleotides and Nucleic Acids*. 2015; 34:7, 449-462.

14. Venter M, Warnich L. In silico promoters: modelling of cis-regulatory context facilitates target predictio. *J Cell Mol Med*. 2008;13(2):270–278.

15. Tatarinova T, Kryshchenko A, Triska M, et al. NPEST: a nonparametric method and a database for transcription start site prediction. *Quant Biol*. 2013;1(4):261–271.

16. Tatarinova TV, Chekalin E, Nikolsky Y, et al. Nucleotide diversity analysis highlights functionally important genomic regions. *Sci Rep*. 2016;6:35730.

17. Khan A, Fornes O, Stigliani A, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res*. 2017;46(D1):D260–D266.

18. Jin J, Tian F, Yang DC, et al. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res*. 2016;45(D1):D1040–D1045.
19. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. *Nucleic Acids Res*. 2008;36.
20. Xia X. Position weight matrix, gibbs sampler, and the associated significance tests in motif characterization and prediction. *Scientifica (Cairo)*. 2012;2012:917540.
21. Venter, M., & Warnich, L. *In silico* promoters: modelling of *cis*-regulatory context facilitates target predictio. *Journal of Cellular and Molecular Medicine*. 2009;*13*(2), 270–278.
22. Tatarinova, T., Kryshchenko, A., Triska, M. et al. NPEST: a nonparametric method and a database for transcription start site prediction. *Quant Biol*. 2013;1:261.
23. Kondrakhin YV, Kel AE, Kolchanov NA, Romashchenko AG, Milanesi L. Eukaryotic promoter recognition by binding sites for transcription factors. *Comput Appl Biosci.* 1995;11(5):477–88.
24. Prestridge DS. Predicting Pol II promoter sequences using transcription factor binding sites. *J Mol Biol*. 1995;249(5):923–32.
25. Weirauch MT, Yang A, Albu M, et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*. 2014;158(6):1431–1443.
26. Li Y, Xiao J, Chen L, Huang X, Cheng Z, Han B, Zhang Q, Wu C. (). Rice Functional Genomics Research: Past Decade and Future. *Cell*. 2018;11(3):359-380.
27. Jiang Y, Cai Z, Xie W, Long T, Yu H, Zhang Q. Rice Functional Genomics Research: Progress and Implications for Crop Genetic Improvement. *Biotechnology advances*. 2011;30(5):1059-1070.
28. Raab JR, Kamakaka RT. Insulators and promoters: closer than we think. *Nat Rev Genet*. 2010;11(6):439–446.
29. Schwartz SH, Silva J, Burstein D, Pupko T, Eyras E, Ast G. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res*. 2008;18(1):88–103.
30. Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res*. 2006;34(14):3955–3967.
31. Claverie JM. Some useful statistical properties of position-weight matrices. *Comput Chem*. 1994 Sep;18(3):287-94.
32. Bhagwat M, Aravind L. PSI-BLAST tutorial. *Methods Mol Biol*. 2007;395:177–186.
33. Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1984;6(6):721–741.
34. Mishra R, Joshi RK, Zhao K. Genome Editing in Rice: Recent Advances, Challenges, and Future Implications. *Front Plant Sci*. 2018;9:1361.
35. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res*. 2004;14(6):1188–1190.
36. Léon-Kloosterziel, K. M., Gil, M. A., Ruijs, G. J., Jacobsen, S. E., Olszewski, N. E., Schwartz, S. H., Zeevaart, J. A. and Koornneef, M. Isolation and characterization of abscisic acid-deficient *Arabidopsis* mutants at two new loci. *The Plant Journal*. 1996;10: 655-661.

37. Okamoto, M. , Tatematsu, K. , Matsui, A. , Morosawa, T. , Ishida, J. , Tanaka, M. , Endo, T. A., Mochizuki, Y. , Toyoda, T. , Kamiya, Y. , Shinozaki, K. , Nambara, E. and Seki, M. Genome-wide analysis of endogenous abscisic acid-mediated transcription in dry and imbibed seeds of Arabidopsis using tiling arrays. *The Plant Journal*. 2010;62:39-51.

38. Feurtado JA, Huang D, Wicki-Stordeur L, Hemstock LE, Potentier MS, et al. The Arabidopsis C2H2 zinc finger Indeterminate Domain1/Enhydrous promotes the transition to germination by regulating light and hormonal signaling during seed maturation. *The Plant Cell* 23. 2011:1772–1794.

39. Giri J, Vij S, Dansana PK, Tyagi AK. Rice A20/AN1 zinc-finger containing stress-associated proteins (SAP1/11) and a receptor-like cytoplasmic kinase (OsRLCK253) interact via A20 zinc-finger and confer abiotic stress tolerance in transgenic Arabidopsis plants. *New Phytol*. 2011;191:721–732.

40. Sorensen, A. , Kröber, S. , Unte, U. S., Huijser, P. , Dekker, K. and Saedler, H. The *Arabidopsis ABORTED MICROSPORES* (*AMS*) gene encodes a MYC class transcription factor. *The Plant Journal*. 2003;33:413-423.

41. Yanhai Yin, Dionne Vafeados, Yi Tao, Shigeo Yoshida, Tadao Asami and Joanne Chory, A New Class of Transcription Factors Mediates Brassinosteroid-Regulated Gene Expression in Arabidopsis, *Cell*. 2005;120(2):(249-259).

42. Ohme-Takagi M, Shinshi H. Ethylene-inducible DNA binding proteins that interact with an ethylene-responsive element. *Plant Cell*. 1995;7(2):173–182.

43. Suzuki, K. , Suzuki, N. , Ohme-Takagi, M. and Shinshi, H. Immediate early induction of mRNAs for ethylene-responsive transcription factors in tobacco leaf strips after cutting. *The Plant Journal*. 1998;15: 657-665.

44. Eyal, Y. , Meller, Y. , Lev-Yadun, S. and Fluhr, R. A basic-type PR-1 promoter directs ethylene responsiveness, vascular and abscission zone-specific expression. *The Plant Journal*. 1993;4:225-234.

45. Fujimoto SY, Ohta M, Usui A, Shinshi H, Ohme-Takagi M. Arabidopsis ethylene-responsive element binding factors act as transcriptional activators or repressors of GCC box-mediated gene expression. *Plant Cell*. 2000;12(3):393–404.

46. Oñate-Sánchez L, Anderson JP, Young J, Singh KB. AtERF14, a member of the ERF family of transcription factors, plays a nonredundant role in plant defense. *Plant Physiol*. 2007;143(1):400–409.

47. Yang, S. , Wang, S. , Liu, X. , Yu, Y. , Yue, L. , Wang, X. and Hao, D. Four divergent *Arabidopsis* ethylene-responsive element-binding factor domains bind to a target DNA motif with a universal CG step core recognition and different flanking bases preference. *The FEBS Journal*. 2009;276:7177-7186.

48. Bolt, S., Zuther, E., Zintl, S., Hincha, D. K., and Schmülling, T. *ERF105* is a transcription factor gene of *Arabidopsis thaliana* required for freezing tolerance and cold acclimation. *Plant, Cell & Environment*. 2017;40:108– 120.

49. Agarwal, P.K., Agarwal, P., Reddy, M.K. et al. *Plant Cell Rep*. 2006;25:1263.

50. Park JM, Park CJ, Lee SB, Ham BK, Shin R, Paek KH. Overexpression of the tobacco Tsi1 gene encoding an EREBP/AP2-type transcription factor enhances resistance against pathogen attack and osmotic stress in tobacco. *Plant Cell*. 2001;13(5):1035–1046.

51. Gu YQ, Wildermuth MC, Chakravarthy S, et al. Tomato transcription factors pti4, pti5, and pti6 activate defense responses when expressed in Arabidopsis. *Plant Cell*. 2002;14(4):817–831.

BIOGRAPHY

Shiva Rawat received his Bachelor of Arts in Biology from the University of Virginia in 2009. He then took the GRE and computer science courses at Northern Virginia Community College in order to prepare for graduate school. Shiva continues his studies at George Mason University, where he is currently finishing up his Masters of Science degree in Bioinformatics and Computational Biology.