Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Journal of Genomes and Exomes

Genome-wide Identification and Characterization of Transcription Factor Binding Motifs of NBS-LRR Genes in Rice and *Arabidopsis*

Gandhimani Ramkumar[1], Maganti S. Madhav[1], Akshaya K. Biswal[2], S.J.S. Rama Devi[1], Kannabiran Sakthivel[3], Madhan K. Mohan[1], B. Umakanth[1], Satendra K. Mangrauthia[1], Raman M. Sundaram[1] and Basavaraj C. Viraktamath[1]

[1]Directorate of Rice Research, Rajendranagar, Hyderabad, India. [2]Plant Breeding and Biotechnology, International Rice Research Institute, Philippines. [3]TNAU-Vegetable Research Station, Palur, India.

ABSTRACT: Nucleotide binding site leucine rich repeat (NBS-LRR) gene encoding proteins are the major biotic stress resistance genes of rice and *Arabidopsis thaliana*. Upstream sequences of 206 entire NBS-LRR genes of *Arabidopsis* and 120 genes of rice were analyzed with three highly reliable motif prediction tools for enhanced accuracy of prediction and characterization of potential transcription factor binding motifs (TFBMs). A total of 36 and 30 novel, strong TFBMs were discovered from NBS-LRR genes of rice and *A. thaliana*, respectively. All the motifs identified in these sequences were analyzed for their positional conservation and the possible motif network associations were also identified. Further, the probability of the presence of motifs in these NBS-LRR genes were validated and statistically tested. Although *Arabidopsis* NBS-LRR sequences showed 76.3% similarity with rice sequences at motif level, the analysis revealed that rice sequences have many unique TFBMs and are more evolved in gene expression mechanisms. The study also provided a list of novel candidate motifs for these genes, which will be a good resource for experimental validation. A novel strategy of prediction of gene expression based on motif arrangement was also demonstrated in this study. The findings of this study, such as the motifs' positional conservation, architecture, etc. offered new biological insights into the role of TFBMs in the regulation of resistance genes.

KEYWORDS: NBS-LRR, transcription factor binding motifs, cis-acting elements, strong motif, motif modules, positional conservation

CITATION: Ramkumar et al. Genome-wide Identification and Characterization of Transcription Factor Binding Motifs of NBS-LRR Genes in Rice and *Arabidopsis*. *Journal of Genomes and Exomes* 2014:3 7–15 doi:10.4137/JGE.S13945.

RECEIVED: December 16, 2013. **RESUBMITTED:** January 27, 2014. **ACCEPTED FOR PUBLICATION:** January 28, 2014.

ACADEMIC EDITOR: Stephen F. Kingsmore, Editor in Chief

TYPE: Original Research

FUNDING: Authors acknowledge the Department of Biotechnology, New Delhi.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: sheshu_24@yahoo.com

Introduction

Biotic stresses, which include diseases and insect pests, greatly affect the crop growth and yield, resulting in declined productivity. These stresses induce up- and down-regulation of different varieties of genes, whose products function not only for stress response but also to impart disease resistance or tolerance. In the signal transduction network, from the perception of stress signals to stress responsive gene expression, various transcription factors (TFs) and cis-acting elements present in the upstream of stress responsive genes function together to cope with pathogen attack for survival and adaptation. More than 1500 genes (>5% of the genome) were predicted to encode for TFs in *Arabidopsis thaliana*.[1] Many of these TFs are involved in various defense pathways and are responsible for the orchestration of pathogen induced gene expression by binding with specific regulatory sites in the promoter region called transcription factor binding motifs (TFBMs) or simply, motifs, which can regulate the spatio-temporal expression of genes.[2] Promoters from co-expressed and orthologous genes may harbor similar TFBMs[3] and those TFBMs represent a particular group of TFs indirectly.[4] Dissection of the promoter content and analysis of the architecture of TFBMs would lead to better understanding of gene regulation at the transcription level in stress responsive genes.[5] Although the methods like

chromatin immunoprecipitation (ChIP), staggered promoter deletions, and DNase I footprinting assays are efficient in the identification of TFBMs, they are time consuming, cumbersome, and also may not cover the entire upstream region.[6,7] In silico methods can predict TFBMs in a wide range of upstream regions of target genes with less effort, time, and cost, making functional analysis of the TFBMs simpler and more efficient. In eukaryotes, transcription related regulatory motifs were repeated many times in the upstream regions[8] and these over represented motifs were correlated with transcriptional activity.[9] However, identification of over represented "true" motifs in eukaryotic genome is still a challenge because of shorter motif size (usually 4–20 bp) in a large genome,[10,11] high degeneracy of nucleotides in the motif sequence (up to 50%),[12] and their wide spread presence over long distance, ie distal parts of the promoter.[13] To overcome these problems, many bioinformatic tools have been developed for the precise identification of motifs. Among them, W-AlignACE,[14] MEME (Multiple EM for Motif Elicitation),[15] and Weeder Web[16] are well known for their reliable and efficient discovery of regulatory motifs. Notably, in a study on the comparative assessment of the performance of various motif discovery tools, Weeder Web and MEME out performed others in many measures.[11] It has also been suggested that use of more than one tool and considering the motifs predicted by complementary tools rather than a single tool would increase the accuracy in predicting the potential motifs.[11] Despite the availability of online tools with a combination of different programs, like Melina II,[17] most of them are not being upgraded and hence, we used these specific programs individually.

Nucleotide binding site plus leucine rich repeat (NBS-LRR) domain containing genes are the most prevalent class of resistance genes in plants.[18] These genes are expected to be orthologous, as they have common targets, ie stress response. Wang et al[7] have attempted to identify the cis-regulatory elements from sorghum and rice from random co-expressed gene groups by using two complementary bioinformatics tools (PhyloCon and FASTCOMPARE). Lengar and Joshi, 2009[8] also identified functional regulatory elements of genes, which involved in developmental stages of *Plasmodium falciparum* by *in silico* analysis. However, there are no reports on the determination of regulatory motifs of orthologous resistant genes and their possible network. Rice and *A. thaliana* not only have high quality genome sequences in the public domain but are the model plant systems for genomics. Surprisingly, motif diversity analysis in respect of resistance gene expression and regulation studies is not yet reported even from these genomes. Studies on motif discovery and their characterization in these two genomes may provide critical information on the regulatory networks underlying gene expression and may also offer clues in understanding the evolutionary status among monocot and dicot plant systems with respect to resistance genes. Therefore, in the present study, an attempt was made to identify novel regulatory motifs and analyze their frequency, distribution pattern, positional preferences, motif network, and grouping of NBS-LRR resistance genes based on motif architecture in rice and *A. thaliana*. In addition to these, the prediction of gene expression pattern was demonstrated based on motif architecture.

Methods

Upstream sequences. The entire NBS-LRR (206) genes and their promoter regions (1000 bp, upstream from transcription start site, TSS) in *A. thaliana*[18] were extracted from NCBI (www.ncbi.nlm.nih.gov) and TAIR (http://www.arabidopsis.org/tools/index.jsp) databases (Supplementary Table 1). From rice, 120 NBS-LRR sequences, which covered the entire genome, were selected and their upstream regions (1000 bp, before TSS) were extracted from RGAP 6.1 (http://rice.plantbiology.msu.edu/) (Supplementary Table 2).

Motif discovery. The motifs were identified from selected upstream sequences of rice and *A. thaliana* using three motif prediction tools (i) MEME,[15] (ii) W-AlignACE,[14] and (iii) Weeder Web.[16] The MEME tool was locally run to discover the motifs with a cut-off *E*-value of ≥1 and a any number of repetitions (anr) level. Simple repetitive nucleotides of the target sequences were removed using Repeat Masker (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) prior to the analysis with MEME. While running W-AlignACE, the number of columns to align was set from 6 to 10 and the program was run twice for each value and hence resulting in 10 runs per sequence. The expected number of sites was set to five with the fractional background GC content set to 0.44 and 0.35, as the average GC content of rice and *A. thaliana* were 44 and 35%, respectively. Only consensus motifs with a MAP score of ≥10 were considered for further analysis. Weeder Web was run locally in large mode to identify motifs with lengths of 6, 8, 10, and 12 nt, with 1, 2, 3, and 4 recombination(s), respectively. Top 10 motifs with high scores and "advised" motifs were stored for further analysis. Motif comparison between rice and *Arabidopsis* was done with Locator tool with 1 bp degeneracy and at least 90% similarity.

Database search. The motifs identified by the three bioinformatics tools discussed above were compared with each other manually. Although the motifs identified by all three bioinformatic tools had higher significance, motifs returned by at least two programs were only considered as strong motifs.[8] The selected strong motifs were checked with TFBM databases viz., PLACE (Plant cis-acting regulatory DNA elements) (http://www.dna.affrc.go.jp/PLACE/signalscan.html)[19] plant CARE (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/),[20] and literatures for further information about these motifs.

Motif diversity analysis, positional specificity, and motif networks. The positional preference of motifs was analyzed by identifying the position and frequency of each strong motif using the Locator tool of Weeder Web.[16] The upstream region was divided into 10 parts with a window size of 100 bp (like 1000–901, 900–801, … 100–1 bp). The higher

rate of appearance of a specific motif in a particular window was considered to indicate that the motif was positionally conserved in that particular window. Motifs with 90% similarity or 1 bp degeneracy were considered as the same motif.

Motif network (diverse motif arrangement in the promoter region in a particular fashion). In order to assess the motif networks in NBS-LRR genes, a binary matrix for these strong motifs based on the presence or absence of motifs in each NBS-LRR sequences of rice and *Arabidopsis* was prepared individually. The binary data was used to construct a dendrogram using NTSYSpc 2.0,[21] which revealed the percentage of motif coordination in gene expression by their position and the closeness between the motifs. If two motifs involve in motif coordination, they are expected to be bound together in most of the sequences and they would get closer position and a high percentage of motif network score in the motif dendrogram. Motifs with overlapping nucleotide sequences were avoided while calculating motif coordination percentage. As genes with the same motif organization and combinations are likely to express for a common cause, an attempt was made to group the NBS-LRR genes based on the architecture of motifs and their combinations.[21]

Validation of strong motifs. In order to distinguish motifs associated the NBS-LRR gene from other common motifs, the strong motifs were validated with two sets of negative control sequences: (i) upstream sequences (1 kb) of 50 housekeeping genes (Supplementary Tables 3 and 4 for rice and *Arabidopsis*, respectively) and (ii) 50 random sequences from all chromosomes (non NBS-LRR genes). Upstream sequences (1 kb) of reported 15 NBS-LRR genes were used as positive control sequences. Presence/proportion of strong motifs in the test sequences have been cross checked with negative as well as with positive control sequences using the Locator tool and their frequency (in percentage) was calculated. If the proportion (percentage) of a particular motif is higher in NBS-LRR promoter sequences than those of negative control sequences, that motif was considered as NBS-LRR associated motif. If the motif was observed only in the NBS-LRR sequences and was absent in the negative control sequences, it was considered as a unique motif.

In order to validate the identified genes clusters in this study, the rice genes groups were compared with gene expression data of the rice annotation project (http://rice.plantbiology.msu.edu/expression.shtml). The rice gene groups were also compared specifically with the transcript data generated through MPSS and SBS libraries made in resistant and susceptible rice cultivars, after *Xanthomonas* and *Magnaporthe oryzae* infections (unpublished, data generated at Dr Wang's Lab, Ohio State University, USA) to check for the expression pattern of R genes.

Statistical analysis. In order to determine the significance of the presence of the motifs obtained, the motifs were analyzed using the *Z*-test. The analysis was carried out by comparing the ratio of occurrence of the motif in the target sequence to the ratio of occurrence of the motif in the two control sequences ie (housekeeping genes and random sequences) using the formula,

$$Z = |p_1 - p_2| / \sqrt{p_1 q_1 / n_1 + p_2 q_2 / n_2}$$

where P is proportion of presence of motif in the target sequences, Q is $1 - P$, which indicates the proportion of motif in the non target sequences, and n is number of sequences tested. The calculated Z values were compared with the standard Z tabulated values, 1.96 ($P < 0.05$) and 2.576 ($P < 0.01$) to find the significance of the motifs identified.

Results and Discussion

Motif search and identification of strong motifs. A total of 113 and 81 strong motifs were identified from NBS-LRR sequences of rice and *Arabidopsis* (Supplementary Tables 5 and 6), respectively. These motifs were found to be distributed over the entire 1 kb upstream region and the motif lengths varied from 5 to 17 bp, while the majority of the motifs (60–65%) were in the range of 6–8 bp and the mean motif lengths was 8.2 ± 2.1 bp for both the genomes. The function of most of the identified strong motifs ie 77 motifs (68.1%) of rice and 51 motifs (63.0%) of *Arabidopsis* were reported earlier (Supplementary Tables 5 and 6, respectively), which verifies the reliability of the motif prediction strategy followed in this study. Hence, the present study led to the identification of 36 and 30 novel, strong motifs in rice and *Arabidopsis,* respectively, which may be the functional TFBMs for unknown TFs and may play a major role in the regulation of genes containing NBS-LRR domain. This study covers the major segment of NBS-LRR associated motifs; for example, Park et al has reported a biotic stress related motif, TTTGAAAACTT, involved in pathogen/environmental/heat stress response (Table 1).[22] Zhang et al also reported motifs, such as TGACCT, involved in pathogen response (Supplementary Table 5),[22] and many other pathogen-defense related motifs were identified in our present study. These novel motifs will be useful resources for experimental validation and can be used to manipulate the gene expression.

Basic and common TFBMs like TATA box, CAAT box, and GATA elements, were not considered, as they are common to most of the genes. Although unique motifs are not expected for every set of co-expressed genes,[8] to identify the motifs that are present in majority of genes (over represented), a thorough scan was made. Interestingly, more than half of the analyzed rice and *Arabidopsis* sequences have showed the presence of five and six over represented motifs, respectively. One such motif (AAAAT), which was reported to be involved in light response,[20] was found to be present in 95 of 120 rice NBS-LRR sequences with 519 hits (number of repetition in sequences). Similarly, another motif CGGAAA, reported to be involved in wound response,[20] was found to have 195 hits

Table 1. Rice and *Arabidopsis* NBS-LRR sequence specific motifs.

S. NO	RICE MOTIF	FUNCTION	REFERENCE	*ARABIDOPSIS* MOTIF	FUNCTION	REFERENCE
1	ACGTGGCTGGTAC	*A. thaliana*: ABA stress response element	24	CCGAACTGCCCT	*A. thaliana*: water stress response element	20
2	GACGTGGGGCCT	*A. thaliana*: ABA stress response element	24	CGGTGGCT	*Oryza sativa*: ABA responsive element	20
3	TCAGACTGCC	*Catharanthus roseus*: jasmonate and elicitor-responsive element	20	GCAGCACCCC		NYR
4	TTTGAAAACTT	Glycine max pathogen/ environmental/heat stress responsive element	25	AGCAGCACCCCA		NYR
5	AAGGGGCGACGC		NYR	AGTACGCGTCCT		NYR
6	AATATTGAAAGA		NYR	TAGTACGCGTCC		NYR
7	CGGTAGGGGG		NYR	GCGAGCTAAGG		NYR
8	ACAACGACGTGG		NYR	GGCGAGCTAAGG		NYR
9	CGTGGGGCCTGG		NYR	CCAGGCTGCT		NYR
10	TCAGATTGCCAT		NYR	CTCCAGGCTGCT		NYR
11	TGGGGCCTGGCG		NYR	TGGTGCTAGGCTGG		NYR
12	CCGTGCAGGCGG		NYR	TATGCAGCATGTCGTAG		NYR
13	CGCGGGCCACCA		NYR	CGAACTGCCCTC		NYR
14	TTCGGGGACCTT		NYR			NYR

Abbreviation: NYR, not yet reported.

in 85 of 120 NBS-LRR sequences, which was followed by GCCCTG (regulatory element associated with GCN4) with 107 hits in 67 sequences (Fig. 1). In *Arabidopsis*, the motif, AAACGAA (involved in elicitor-responsive element) was present in 166 sequences with 436 hits, GAAGAC (heat shock element) was present in 164 sequences, and had the highest hits of 487. Another motif, CGCCAA (reported as enhancer site) was present in 141 sequences with 294 hits (Fig. 2). As these motifs were over represented in many sequences with high frequency, these motifs are highly significant and expected to be the major TFBMs in most of the NBS-LRR containing genes.[23] The web logo of these over represented motifs of rice and *Arabidopsis* are given in Figures 1 and 2, respectively. However, the roles of some of the over represented motifs are not known and these could be the novel TFBMs. Top 20 over represented motifs occurring in higher frequency (based on the presence of motifs in the number of sequences) are given in Figures 3 and 4 for rice and *A. thaliana*, respectively. To identify the NBS-LRR sequence specific motifs for this class of R genes, the strong motifs were validated with two sets of negative control sequences; promoter sequences of housekeeping genes, and random sequences (non NBS-LRR sequences). This experiment revealed that majority of the motifs of rice and *Arabidopsis* had higher proportion in NBS-LRR sequences than the negative control sequences and comparison of strong motifs to the negative control motifs indicated that no motif was false negative. Among those, 14 and 13 motifs of rice and *Arabidopsis*, respectively, were unique to NBS-LRR genes (motifs were present only in NBS-LRR sequences and absent in both the negative controls), (Table 1). Interestingly, among those 14 motifs of rice, the function of four motifs has been reported to be involved in pathogen reaction either directly or indirectly[24,25] and the functions of the remaining 10 motifs are not yet reported. In *Arabidopsis*, among the 13 NBS-LRR associated motifs, the functions of two motifs (related to water stress and abscisic acid (ABA) response elements) have been reported, while the remaining 11 are novel motifs. These unique motifs of NBS-LRR genes could be considered as candidate motifs and can be experimentally validated for their role in pathogen reaction. Statistical analysis with *Z*-test also revealed the presence of two motifs (CGGGGACC and TTTGAAAACTT), which were significantly present among NBS-LRR sequences compared to the other two control sequences. Contrary to rice, *Arabidopsis* NBS-LRR sequences had 31 statistically significant motifs when compared to the two control sequences.

Extensive studies reported that regulatory elements are accumulated in the regulation region up to –500 bp commonly.[26] The results of Segal et al indicated that motifs were concentrated at the region near the TSS and the number of motifs decreased significantly after –250 bp.[27] The results of Zheng et al also suggested that the motifs were accumulated

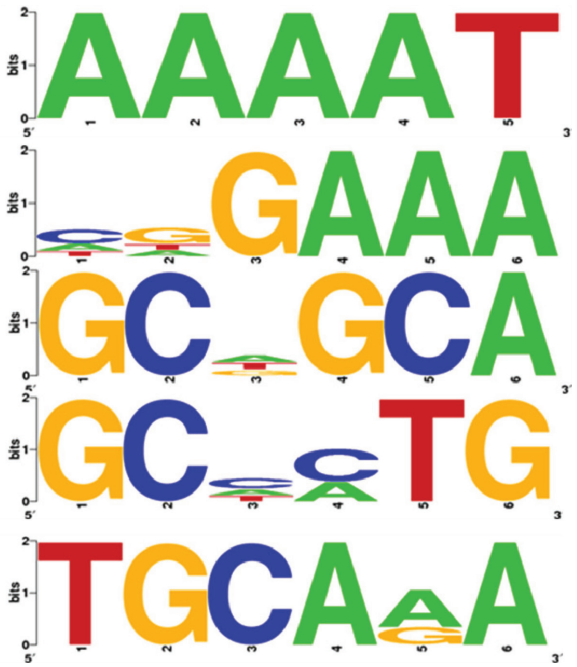

Figure 1. Among the identified 113 motifs from rice NBS-LRR promoter sequences, five motifs were present in majority (more than 50%) of the analyzed sequences; this high frequency of motifs may be significant and expected to be major TFBMs to most of the NBS-LRR sequences. The picture represents web logos of those high frequency strong motifs of rice.

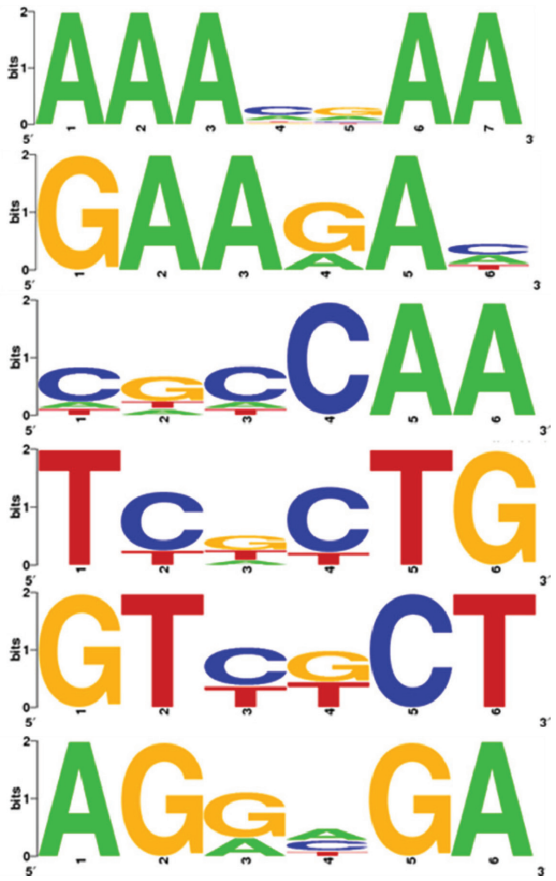

Figure 2. Among the identified 81 motifs from *Arabidopsis* NBS-LRR promoter sequence, six motifs were present in majority (more than 50%) of the analyzed sequences; this high frequency of motifs may be significant and expected to be major TFBMs to most of the NBS-LRR sequences. The picture represents the web logo of the high frequency strong motifs of *Arabidopsis*.

within the −400 bp region than the distal part of the regulatory region.[28] Moreover, the true motif prediction accuracy will be reduced if the length of the sequence is increased.[29] Hence, to balance the motif prediction quantity and the accuracy, we analyzed 1 kb upstream region in this study. It is noteworthy that many other studies also analyzed 1 Kb or less than 1 kb upstream region for this purpose.[30,26]

We believe that the present study could cover major portion of the motifs of this group. For example, Park et al reported biotic stress related motif, TTTGAAAACTT, involved in pathogen/environmental/heat stress response (Table 1);[25] Zhang et al also reported motifs, such as TGACCT, involved in pathogen response (Supplementary Table 5),[22] and many other pathogen-defense related motifs were also identified (Table 1) in our present study.[20] Some exclusion of motifs could be because of the stringent parameters followed in our study to increase the motif prediction accuracy and because this study did not include the non NBS-LRR biotic stress related genes for analysis. However, we strongly believe that adding the strong and novel motifs, identified in this study, to the known/already identified motifs, will be helpful to the necessary motif validation study and to enhance our knowledge in this resistance genes expression segment.

Positional conservation of motifs. It has been shown earlier that positional conservation of motifs has high correlation with transcriptional machinery.[23] Among the identified

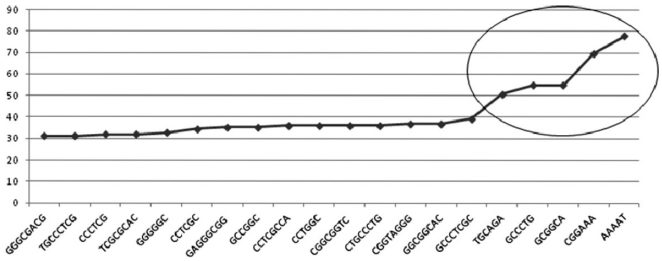

Figure 3. Frequency of the top 20 motifs (based on their frequency) in the rice NBS-LRR gene promoter sequences. The motif, AAAAT, which had the highest frequency, was present in 95 out of 120 sequences, while the motif, GGGCGACG was present in 38 out of 120 sequences (least among the top 20 motifs). In the figure, *Y*-axis represents the percentage of frequency.

strong motifs, 65.5% (74 of 113) of rice and 70% (57 of 81) of *Arabidopsis*, exhibited positional preferences at different windows of the upstream region. A majority of the positionally conserved strong motifs were located within the first 500 bp in the tested sequences (Fig. 5), which suggest that most of

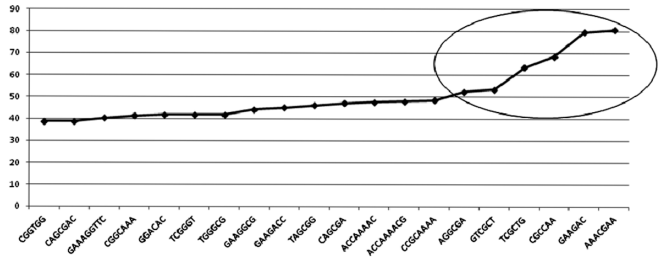

Figure 4. Frequency of the top 20 motifs (based on their frequency) in the *Arabidopsis* NBS-LRR gene promoter sequences. The motif AAACGAA, which had the highest frequency, was present in 166 out of 206 sequences, while the motif CGGTGG was present in 80 out of 206 sequences (least among top 20 motifs). In the figure, the *Y*-axis represents the percentage of frequency.

the tested R genes may have the minimal promoter length of ~500 bp from the TSS. The motifs involved in defense responses such as GCGGCA and GGGGGC (involved in elicitation; wounding, and pathogen response); CCGCCA, GAGGGCGG (jasmonate- and elicitor-responsive element); CGCCCC, CAGCGAC (ABA stress response element); and AGCGACAG in cold stress were conserved in the proximal promoter region (–1 to –200 bp) in the tested sequences of rice and *Arabidopsis*. This was also in close agreement with the previous reports, where it has been demonstrated in upstream of the putative yeast genes by Brazma et al.[31] These motifs might be significant in the defense response, as these motifs are present in the proximal promoter region and having pathogen resistance related function. Among the 36 and 30 novel motifs of rice and *Arabidopsis*, 23 and 15 motifs displayed positional conservation at different windows of the promoter region in both rice and *Arabidopsis* (Supplementary Table 5 and 6), respectively. The observed positional conservation of the motifs indicated the significance of the strong motifs and suggested that these motifs might have regulatory roles.[32,26] The large proportion of positional conserved

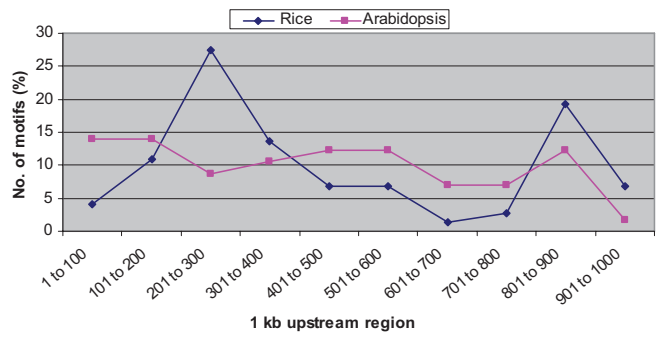

Figure 5. Comparison of strong motifs of rice and *Arabidopsis* in terms of distribution in promoter region, which indicated that rice had highest number of motif conservation at –201–300 window, while the *Arabidopsis* had highest number of motif conservation at –1–100 as well as –101–200 windows.

motifs also suggested that promoter conservation might have occurred during the process of evolution.

Motif network. The regulatory function of the motifs in specific motif combination have to be studied in detail for better understanding of the regulatory network of the genes.[33,34] It was reported that different combinations of less numbers of motifs regulate gene expression in many organisms rather than a large variety of TFBMs.[8,35] Expression of many genes in a cell rely on the combinatorial control of several TFs[36,37] and in order to understand those cellular processes better, motif combinations also have to be analyzed along with novel motif identification. Occurrence of motif network in promoters of late embryogenesis abundant genes in rice and their use in finding the co-expressed/complementary genes by analyzing their motifs pattern was demonstrated efficiently.[34,38] The genes that have similar kind of motif arrangements are generally expected to be functionally related and may co-express with a common trigger.[38] Two motif regulatory modules consisting of ACGTGGCTGGTAC (ABA stress responsive element) and TGGGGCCTGGCG (unknown function) were observed to be highly coordinated in motif combination among all the tested rice NBS-LRR sequences and thus exhibited 100% association between them (Supplementary Fig. 1). Another rice motif regulatory module consisting of three motifs, known to be involved in defense response viz., CGGCGC (ABA response element), GCCGGC (wounding and pathogen response), and GGGGGCGG (repressor element) were also observed with 68% motif coordination in the sequences, wherever these motifs exist.

Similarly, four motif regulatory modules, where some members were known to play pathogen response role viz., GCGGCA (pathogen response), GCCCTG (regulatory element associated with GCN4), CGGAAA (wound response element), and AAAAT (light responsive element) were observed with 63% motif coordination in the sequences. Interestingly, one of the over represented motif of rice also formed a module with the other motifs involved in the defense response, which clearly indicated the existence of higher motif coordination among the motifs (Supplementary Fig. 1). In *Arabidopsis*, motif modules consisting of five and two motifs were observed with 50% coordination. Other combinational motif modules of *Arabidopsis* sequences are illustrated in Supplementary Figure 2. It is observed that, the possibility of forming motifs into modules and their coordination percentages was inversely related; if the number of motifs in a module increases, the coordination percentage will be decreased. Although it cannot be expected that all the NBS-LRR genes will express at same time for a particular defense response, it is possible that a set of NBS-LRR genes may co-express for a common cause and that can be predicted by analyzing the combinations of motifs. Based on motif module presence, 21 NBS-LRR genes out of 113 genes had three motif modules (CGGCGC, GCCGGC, and GGGGGCGG) (Supplementary Fig. 3) and these genes might be co-expressed by

common external stimulus. Similarly, seven NBS-LRR rice genes also possessed four motif module coordination (GCG-GCA, GCCCTG, CGGAAA, and AAAAT), which may co-express (Supplementary Fig. 4).

To validate the identified gene clusters in this study, these clusters were compared with gene expression data of the rice annotation project (http://rice.plantbiology.msu.edu/expression.shtml) and also biotic stress transcriptome data. Interestingly, many gene clusters showed a similar kind of expression pattern. For instance, the two identified genes in a cluster, LOC_Os02g30150 and LOC_Os12g10180 showed a similar kind of expression pattern; ie they expressed during the cold stress. Most of the above mentioned cluster containing 21 genes expressed in mature pollen. Hence, this study proved that the strategy of gene clustering based on its motif arrangement, correlated to the gene expression. Comparison of biotic stress transcriptome data also revealed that 11 of the rice genes showed expression in *M. oryzae* infected rice leaves (two in very early stage of infection, ie three hours after infection [Os11g11960, Os11g45750], three in early stage of infection, ie six hours after infection [OsOs11g42040, Os11g12260, Os11g12000], two in middle stage of infection, ie 12 hours after infection [Os01g71106, Os11g45190], and four in late, ie 48 hours after infection [Os12g13550, Os06g17970, Os12g28100, Os12g10180] infections) and 10 of the rice genes expressed in *Xanthomonas* infected leaves (one in very early [Os09g34150], five in early [Os12g17430, Os04g52970, Os11g29520, Os07g29820, Os11g15700], four [Os02g18070, Os11g13940, Os08g10260, Os11g45060] in middle stage of infection) (Supplementary Fig 3). This further suggested that the adopted strategy of the present study would facilitate the prediction of gene expression based on motif arrangement. Hence, the demonstrated strategy in the present study could be useful to predict the expression pattern of the genes and to identify co-expressive genes not only related to stress but also related to different groups like non-stress and different stages of plant growth, etc.

Comparative analysis of rice and *Arabidopsis*. Consistent with the estimated timing of their divergence, monocot and eudicot plants share common TF[39] gene families, and the number of genes in each family is similar between rice and *Arabidopsis*.[40] The observed conservation of TF genes indicated that many TFs and TFBMs existed prior to the divergence of monocots and eudicots and they might have evolved from a common ancestor.[39] Similarly, this study revealed that the motifs of *Arabidopsis* and rice NBS-LRR genes also shared 76.3% similarity. Rice motifs show higher GC content (78.4%) than *Arabidopsis* motifs (64.8%), which might be because of the higher GC content of the rice (44%) than the *Arabidopsis* (35%).[41] However, rice NBS-LRR sequences have more specific motifs as compared to *Arabidopsis* (Fig. 6), which indicate that rice R genes have more evolved architecture of gene regulation compared to *A. thaliana*. On the contrary, motifs of *A. thaliana* were general

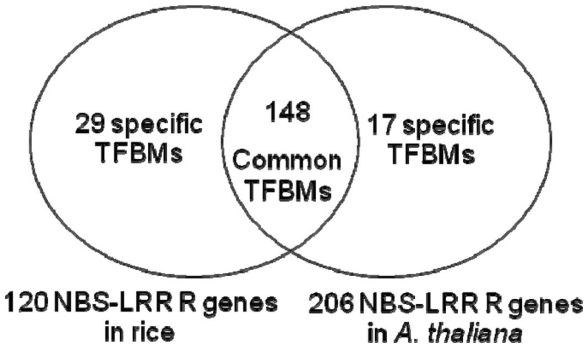

Figure 6. Comparative analysis of rice and *Arabidopsis* at the motif level. Although they shared 148 (as identified by locator program) out of 194 motifs (113 of rice identified from 120 NBS-LRR sequences and 81 of *Arabidopsis* identified from 206 NBS-LRR sequences), rice had more unique motifs (29) than *Arabidopsis*, which had only 17 unique motifs.

and found upstream in a vast number of genes. In regards to the conservation of motifs also, *A. thaliana* motifs have showed higher conservation than rice motifs, which indicated a lower level of polymorphism in the upstream region of *Arabidopsis*. In addition to that, rice motifs have shown higher motif network than *A. thaliana*. Based on these observations, rice NBS-LRR genes could be more evolved in terms of regulatory mechanism than *A. thaliana*.

Conclusion

An attempt was made to characterize promoter regions of NBS-LRR genes and identify the strong TFBMs using popular bioinformatics tools. The frequency and positional conservation of the motifs were determined, which helped in understanding the architecture of the corresponding promoters. We also determined the cis-regulatory modules for these genes using a novel method motif network dendrogram. This information helped in finding the genes that may be co-expressed/complementary to each other. Hence, the present study has resulted in successful identification of unique motifs for the rice and *Arabidopsis* NBS-LRR genes and revealed the cis-regulatory modules that may be playing a role in the cascade of gene expression. Alternatively, the demonstrated method can be used to predict the gene expression pattern based on the motif arrangement. This may help in predicting the expression of any set of genes, based on their motif arrangement pattern. The comparison of the architecture of promoters of rice with *Arabidopsis* at the motif level gave clues in understanding the evolutionary development and the gene regulation. The identified novel and over represented motifs will be a good resource for functional validation, which will lead to a better understanding of the process of resistance gene regulation.

Acknowledgments

The authors sincerely thank Dr G. L. Wang, OSU for sharing the biotic stress transcriptome data. The authors also thank

Dr P. Rajendrakumar, Directorate of Sorghum Research, Hyderabad, for critically reviewing the manuscript and the anonymous reviewers for their important suggestions for improving the manuscript.

Author Contributions

MSM conceived and designed the experiments. GR, AKB, SJSRD, KS, MKM, and BU analyzed the data. GR wrote the first draft of the manuscript. GR and SJSRD contributed to the writing of the manuscript. SKM, RMS, and BCV agree with manuscript results and conclusions. MSM, GR, SJSRD jointly developed the structure and arguments for the paper. MSM made critical revisions and approved the final version. All authors reviewed and approved the final manuscript.

DISCLOSURES AND ETHICS

As a requirement of publication the authors have provided signed confirmation of their compliance with ethical and legal obligations including but not limited to compliance with ICMJE authorship and competing interests guidelines, that the article is neither under consideration for publication nor published elsewhere, of their compliance with legal and ethical guidelines concerning human and animal research participants (if applicable), and that permission has been obtained for reproduction of any copyrighted material. This article was subject to blind, independent, expert peer review. The reviewers reported no competing interests.

Supplementary Data

Supplementary figure 1. Combinational network of rice motifs with each other. Significant combinational motifs have been marked with circle.

Supplementary figure 2. Combinational network of *A. thaliana* motifs with each other. Significant combinational motifs have been marked with circle.

Supplementary figure 3. Rice gene clusters based on different architecture of motifs. (Gene cluster based on CGGCGC, GCCGGC and GGGGGCGG motifs (Genes, which contains all three motifs) were marked with circle).

Supplementary figure 4. Rice gene clusters based on different architecture of four motifs. (Gene cluster based on four motifs GCGGCA, GCCCTG, CGGAAA and AAAAT were marked with circle).

Supplementary table 1. 206 NBS-LRR gene sequences of *A. thaliana*.

Supplementary table 2. 120 NBS-LRR sequences of rice.

Supplementary table 3. Fifty housekeeping rice genes list used for strong motif validation.

Supplementary table 4. Fifty housekeeping *Arabidopsis* genes list used for motif validation.

Supplementary table 5. List of strong rice motifs, identified by bioinformatics tools with their positional preferences and function.

Supplementary table 6. List of strong *A. thaliana* motifs, identified by bioinformatics tools with their positional preference and function.

REFERENCES

1. Riechmann JL, Heard J, Martin G, et al. *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science.* 2000;290(5499): 2105–2110.
2. Tjian R, Maniatis T. Transcriptional activation: a complex puzzle with few easy pieces. *Cell.* 1994;77(1):5–8.
3. Lyons TJ, Gasch AP, Gaither LA, Botstein D, Brown PO, Eide DJ. Genome-wide characterization of the Zap 1p zinc-responsive regulation in yeast. *Proc Natl Acad Sci U S A.* 2000;97(14):7957–7962.
4. Thakurta DG. Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Res.* 2006;34(12):3585–3598.
5. Werner T, Fessele S, Maier H, Nelson PJ. Computer modeling of promoter organization as a tool to study transcriptional coregulation. *FASEB J.* 2003;17(10): 1228–1237.
6. Kuo MH, Allis CD. In vivo cross-linking and immunoprecipitation for studying dynamic Protein: DNA associations in a chromatin environment. *Methods.* 1999;19(3):425–433.
7. Wang X, Haberer G, Mayer KF. Discovery of cis-elements between sorghum and rice using co-expression and evolutionary conservation. *BMC Genomics.* 2009;10:284.
8. Lengar P, Joshi NV. Identification of putative regulatory motifs in the upstream regions of co-expressed functional groups of genes in *Plasmodium falciparum*. *Genomics.* 2009;10:1–21.
9. Zhang J, Hu J, Shi XF, Cao H, Liu WB. Detection of potential positive regulatory motifs of transcription in yeast introns by comparative analysis of oligonucleotide frequencies. *Comput Biol Chem.* 2003;27(4–5):497–506.
10. Helden VJ, Andre B, Vides CJ. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol.* 1998;281(5):827–842.
11. Tompa M, Li N, Bailey TL, et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol.* 2005;23(1):137–144.
12. Poluliakh N, Nakai K. Extraction of biological motifs by Gibbs Sampler from the promoters of *Homo sapiens, Saccharomyces cerevisiae* and *Bacillus subtilis*. *Genome Inf.* 2003;14:406–407.
13. Caselle M, Cunto FD, Provero P. Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes. *BMC Bioinf.* 2002;3:7.
14. Chen X, Guo L, Fan Z, Jiang T. W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data. *Bioinformatics.* 2008;24(9):1121–1128.
15. Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 2006;34:W369–W373.
16. Pavesi G, Mereghetti P, Mauri G, Pesole G. Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.* 2004;32:W199–W203.
17. Okumura T, Makiguchi H, Makita Y, Yamashita R, Nakai K. Melina II: a web tool for comparisons among several predictive algorithms to find potential motifs from promoter regions. *Nucleic Acids Res.* 2007;35:W227–W231.
18. Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell.* 2003;15(4): 809–834.
19. Higo K, Ugawa Y, Iwamoto M, Korenaga T. Plant cis-acting regulatory DNA elements PLACE database: 1999. *Nucleic Acids Res.* 1999;27(1):297–300.
20. Lescot M, Dehais P, Thijs G, et al. Plant CARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res.* 2002;30(1):325–27.
21. Rohlf FJ. *NTSYS-PC. Numerical Taxonomy and Multivariate Analysis Systems, Version 2.0.* Setauket: Exeter software; 1997.
22. Zhang ZL, Xie Z, Zou X, Casaretto J, Ho TH, Shen QJ. A rice *WRKY* gene encodes a transcriptional repressor of the gibberellin signaling pathway in aleurone cells. *Plant Physiol.* 2004;134(4):1500–1513.
23. Casimiro AC, Vinga S, Freitas AT, Oliveira AL. An analysis of the positional distribution of DNA motifs in promoter regions and its biological relevance. *BMC Bioinf.* 2008;9:89.
24. Choi HI, Hong J, Ha J, Kang J, Kim SY. ABFs, a family of ABA-responsive element binding factors. *J Biol Chem.* 2000;275(3):1723–1730.
25. Park HC, Kim ML, Kang YH, et al. Pathogen and NaCl-induced expression of the SCaM-4 promoter is mediated in part by a GT-1 box that interacts with a GT-1-like transcription factor. *Plant Physiol.* 2004;135(4):2150–2161.
26. Mohanty B, Krishnan SP, Swarup S, Bajic VB. Detection and preliminary analysis of motifs in promoters of anaerobically induced genes of different plant species. *Ann Bot.* 2005;96(4):669–681.
27. Segal L, Lapidot M, Solan Z, Ruppin E, Pilpel Y, Horn D. Nucleotide variation of regulatory motifs may lead to distinct expression patterns. *Bioinformatics.* 2007; 23(13):i440–i449.
28. Zheng J, Wu J, Sun Z. An approach to identify over-represented cis-elements in related sequences. *Nucleic Acids Res.* 2003;31(7):1995–2005.

29. Hu J, Li B, Kihara D. Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res.* 2005;33(15):4899–4913.
30. Mihara M, Itoh T, Izawa T. In silico identification of short nucleotide sequences associated with gene expression of pollen development in rice. *Plant Cell Physiol.* 2008;49(10):1451–1464.
31. Brazma A, Jonassen I, Vilo J, Ukkonen E. Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.* 1998;8(11):1202–1215.
32. Berendzen KW, Stuber K, Harter K, Wanke D. Cis-motifs upstream of the transcription and translation initiation sites are effectively revealed by their positional disequilibrium in eukaryote genomes using frequency distribution curves. *BMC Bioinf.* 2006;7:522.
33. Zhou Q, Wong WH. Cis module: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci U S A.* 2004;101(33): 12114–12119.
34. Lindlof A, Brautigam M, Chawade A, Olsson O, Olsson B. In silico analysis of promoter regions from cold-induced genes in rice *Oryza sativa*. L and *Arabidopsis thaliana* reveals the importance of combinatorial control. *Bioinformatics.* 2009; 25(11):1345–1348.
35. Noort V, Huynen MA. Combinatorial gene regulation in *Plasmodium falciparum*. *Trends Genet.* 2006;22(2):73–78.
36. Remenyi A, Scholer HR, Wilmanns M. Combinatorial control of gene expression. *Nat Struct Mol Biol.* 2004;11(9):812–815.
37. Singh KB. Transcriptional regulation in plants: the importance of combinatorial control. *Plant Physiol.* 1998;118(4):1111–1120.
38. Meier S, Gehring C, MacPherson CR, et al. The promoter signatures in rice LEA genes can be used to build a co-expressing lea gene network. *Rice.* 2008;1: 177–187.
39. Xiong Y, Liu T, Tian C, Sun S, Li J, Chen M. Transcription factors in rice: a genome wide comparative analysis between monocots and eudicots. *Plant Mol Biol.* 2005;59(1):191–203.
40. Rensink WA, Buell CR. *Arabidopsis* to rice. Applying knowledge from a weed to enhance our understanding of a crop species. *Plant Physiol.* 2004;135(2):622–629.
41. Jander G, Baerson SR, Hudak JA, Gonzalez KA, Gruys KJ, Last RL. Ethylmethanesulfonate saturation mutagenesis in *Arabidopsis* to determine frequency of herbicide resistance. *Plant Physiol.* 2003;131(1):139–146.