Motif content comparison between monocot and dicot species

Matyas Cserhati

Department of Genetics, Cell Biology, and Anatomy, 985805 Nebraska Medical Center, Omaha, NE, 68198-5805, United States

A R T I C L E   I N F O

Article history:
Received 8 October 2014
Received in revised form 29 December 2014
Accepted 30 December 2014
Available online 17 January 2015

Keywords:
Monocot
Dicot
Genome
Motif
Promoter

A B S T R A C T

While a number of DNA sequence motifs have been functionally characterized, the full repertoire of motifs in an organism (the motifome) is yet to be characterized. The present study wishes to widen the scope of motif content analysis in different monocot and dicot species that include both rice species, Brachypodium, corn, wheat as monocots and Arabidopsis, Lotus japonica, Medicago truncatula, and Populus tremula as dicots. All possible existing motifs were analyzed in different regions of genomes such as were found in different sets of sequences in these species: the whole genome, core proximal and distal promoters, 5′ and 3′ UTRs, and the 1st introns. Due to the increased number of species involved in this study compared to previous works, species relationships were analyzed based on the similarity of common motif content. Certain secondary structure elements were inferred in the genomes of these species as well as new unknown motifs. The distribution of 20 motifs common to the studied species were found to have a significantly larger occurrence within the promoters and 3′ UTRs of genes, both being regulatory regions. Motifs common to the promoter regions of japonica rice, Brachypodium, and corn were also found in a number of orthologous and paralogous genes. Some of our motifs were found to be complementary to miRNA elements in Brachypodium distachyon and japonica rice.

© 2015 The Author. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Background

A motif is a conserved or frequently occurring sequence of defined length, usually 4–10 base pairs in the case of DNA sequences. Motifs can be found in DNA/RNA or protein sequences, where each motif is typically associated with certain biological function(s). While a number of DNA sequence motifs have been functionally characterized, the full repertoire of motifs in an organism (referred to as the motifome) is yet to be characterized. In this study, we focus on motifs in plants. The total motif content of motifs in an organism (referred to as the motifome) is yet to be characterized. The present study wishes to widen the scope of motif content analysis in different monocot and dicot species that include both rice species, Brachypodium, corn, wheat as monocots and Arabidopsis, Lotus japonica, Medicago truncatula, and Populus tremula as dicots. All possible existing motifs were analyzed in different regions of genomes such as were found in different sets of sequences in these species: the whole genome, core proximal and distal promoters, 5′ and 3′ UTRs, and the 1st introns. The motifs that we found were matched with experimentally validated regulatory motifs in the Plant Cis-acting Regulatory Elements (PLACE) database [12]. The motifs in this database are characterized by a PLACE ID and a representative sequence, and are also cross-linked to papers describing these motifs in greater detail.

We published similar work on rice in a previous paper [8], and the methodology to find and determine the statistical significance of motifs used in this study is similar to that in previous works [9, 10], therefore we refer the reader to these specific references. Furthermore, what makes the present study significant is that it also makes it possible to compare motif content between more or less related species as well as two different groups of plants, monocots and size of genome) as well as a link or reference to their genomic information is provided in Tables 1 and 2.

The present analysis endeavors to draw up a catalogue of all possible existing octamer motifs (48 = 65,536 in total) in the genomes of the above-named 9 species, as well as determining their statistical significance. In this study these motifs serve as regulatory signals or transcription factor binding sites in promoters, 1st introns, or UTR sequences. Octamers were studied in our previous study of O. sativa japonica [8] and in [9] of Arabidopsis. Motifs that are 8 bp long are long enough to be both diverse enough and statistically significant at the same time. Motifs can be found in different regions of genomes such as were found in 1–7 different sets of sequences, according to its availability in these species: the whole genome, core promoters (250 or 300 bp), proximal promoters (1000 bp), and distal promoters (3000 bp), 5′ and 3′ UTRs, and the 1st introns. The motifs that we found were matched with experimentally validated regulatory motifs in the Plant Cis-acting Regulatory Elements (PLACE) database [12]. The motifs in this database are characterized by a PLACE ID and a representative sequence, and are also cross-linked to papers describing these motifs in greater detail.

We published similar work on rice in a previous paper [8], and the methodology to find and determine the statistical significance of motifs used in this study is similar to that in previous works [9, 10], therefore we refer the reader to these specific references. Furthermore, what makes the present study significant is that it also makes it possible to compare motif content between more or less related species as well as two different groups of plants, monocots and size of genome) as well as a link or reference to their genomic information is provided in Tables 1 and 2.

The present analysis endeavors to draw up a catalogue of all possible existing octamer motifs (48 = 65,536 in total) in the genomes of the above-named 9 species, as well as determining their statistical significance. In this study these motifs serve as regulatory signals or transcription factor binding sites in promoters, 1st introns, or UTR sequences. Octamers were studied in our previous study of O. sativa japonica [8] and in [9] of Arabidopsis. Motifs that are 8 bp long are long enough to be both diverse enough and statistically significant at the same time. Motifs can be found in different regions of genomes such as were found in 1–7 different sets of sequences, according to its availability in these species: the whole genome, core promoters (250 or 300 bp), proximal promoters (1000 bp), and distal promoters (3000 bp), 5′ and 3′ UTRs, and the 1st introns. The motifs that we found were matched with experimentally validated regulatory motifs in the Plant Cis-acting Regulatory Elements (PLACE) database [12]. The motifs in this database are characterized by a PLACE ID and a representative sequence, and are also cross-linked to papers describing these motifs in greater detail.

We published similar work on rice in a previous paper [8], and the methodology to find and determine the statistical significance of motifs used in this study is similar to that in previous works [9, 10], therefore we refer the reader to these specific references. Furthermore, what makes the present study significant is that it also makes it possible to compare motif content between more or less related species as well as two different groups of plants, monocots and size of genome) as well as a link or reference to their genomic information is provided in Tables 1 and 2.

The present analysis endeavors to draw up a catalogue of all possible existing octamer motifs (48 = 65,536 in total) in the genomes of the above-named 9 species, as well as determining their statistical significance. In this study these motifs serve as regulatory signals or transcription factor binding sites in promoters, 1st introns, or UTR sequences. Octamers were studied in our previous study of O. sativa japonica [8] and in [9] of Arabidopsis. Motifs that are 8 bp long are long enough to be both diverse enough and statistically significant at the same time. Motifs can be found in different regions of genomes such as were found in 1–7 different sets of sequences, according to its availability in these species: the whole genome, core promoters (250 or 300 bp), proximal promoters (1000 bp), and distal promoters (3000 bp), 5′ and 3′ UTRs, and the 1st introns. The motifs that we found were matched with experimentally validated regulatory motifs in the Plant Cis-acting Regulatory Elements (PLACE) database [12]. The motifs in this database are characterized by a PLACE ID and a representative sequence, and are also cross-linked to papers describing these motifs in greater detail.

We published similar work on rice in a previous paper [8], and the methodology to find and determine the statistical significance of motifs used in this study is similar to that in previous works [9, 10], therefore we refer the reader to these specific references. Furthermore, what makes the present study significant is that it also makes it possible to compare motif content between more or less related species as well as two different groups of plants, monocots

E-mail address: matyas.cserhati@unmc.edu.
and dicots, making it possible to draw further insights from a wider variety of cross-species comparisons.

**Results and discussion**

**Principle of investigation**

According to the statistical measure, the total occurrence of all possible octamer motifs (65,536 in number) were enumerated in the genomes of the 5 named monocot and the 4 named dicot species. Octamers in the appropriate whole genomes, 5′ UTR, 3′ UTR, 1st introns, core, proximal, and distal promoters were analyzed in some of the monocot species and A. thaliana. Only motifs not containing ambiguous IUPAC symbols (M, R, W, S, Y, K, B, D, H, V, or N) were retained. According to the algorithm, the statistical significance value of the motif occurs in, and ES is the expected number of sequences that the motif occurs in, and ES is the expected number of sequences that the motif is expected to occur in according to the background base distribution of the different species. The genome sequences were masked using the RepeatMasker program prior to analysis to exclude repetitive sequences which might skew the results.

The detailed method of the algorithm is described in previous works [8–10]. However, a short description will be given in the Materials and methods section. All 7 sequence sets were analyzed in Arabidopsis thaliana. Only motifs not containing ambiguous IUPAC symbols (M, R, W, S, Y, K, B, D, H, V, or N) were retained. According to the algorithm, the statistical significance value of the motif occurs in, and ES is the expected number of sequences that the motif occurs in, and ES is the expected number of sequences that the motif is expected to occur in according to the background base distribution of the different species. The genome sequences were masked using the RepeatMasker program prior to analysis to exclude repetitive sequences which might skew the results.

The top 25 octamer motifs found in the monocot and dicot species are listed in Supplementary Tables 1 and 2 respectively; while the entire list of motifs with their significance scores can be found in the Supplementary Excel files for each species (Supplementary file 1 – O. sativa japonica, Suppl. file 2 – O. sativa indica, Suppl. file 3 – Brachypodium, Suppl. file 4 – Z. mays, Suppl. file 5 – Triticum, Suppl. file 6 – dicot genomes, Suppl. file 7 – Arabidopsis compared to monocots). The intersections of octamer motifs between different combinations of monocot and dicot species can be seen in the Venn diagrams in Fig. 1a and b. For all species, the top 100 motifs were further analyzed for their functional significance. The top 100 whole genome motifs were also checked against the PLACE database [12] to see whether they matched any experimentally verified regulatory motifs. The PLACE database [6] is a well-known plant motif database. Each motif is characterized by a PLACE id as well as a functional description as well as Pubmed ID’s which link to papers describing the given motif. The functional definitions of each PLACE id can be found in the supplementary Excel file “PLACE_id_functional_dictionary.xlsx”. Z. mays had the least number of motifs in common with the other 4 species, with 65 of its top 100 high scoring motifs being unique only to itself. Triticum came in a close second with 44 motifs distinct to itself. This is no surprise as Z. mays belongs to a completely different clade (PACC clade) as do the two Oryza species and Brachypodium [13].

Overall, 20 motifs were common to all monocot species, while 41 motifs were common to the dicot species. Of these, 15 were common to all monocot and all dicot species, making them the general plant motif candidates. These motifs can be seen in Table 3. In Table 4 we can see the number of top 100 genomic motifs shared by each of the monocot and dicot species shared between 1, 2, 3, 4, and 5 other species. We can see again that Z. mays has a relatively low number of motifs shared by any number of other species, whereas O. sativa japonica and indica and Brachypodium have 71, 73, and 70 motifs shared by at least three species, 64 motifs common to all three of these species. 98 motifs are common to both rice species, which is also significant.

What is interesting about these motifs is that a large number of them form reverse complementary pairs (AAAAAAAA|TTTTTTTT, AAAAGAAA|TTTCTTTT, AAAGAAAA|TTTTCTTT, AAATAAAA|TTTTTTTT). These might possibly form parts of the secondary stem-loop structure in microRNAs. Indeed, the

| Table 1 | Available data sets for the studied species. |
|---------|---------------------------------------------|
| Species | genome | Core promoters | Proximal promoters | Distal promoters | 1st introns | 5′ UTRs | 3′ UTRs |
| Monocots | | | | | | |
| Oryza sativa japonica | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Oryza sativa indica | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Brachypodium distachyon | 1 | 1 | 1 | X | X | X | 1 |
| Triticum aestivum | 1 | X | X | X | X | X | X |
| Dicots | | | | | | |
| Arabidopsis thaliana | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Lotus japonica | 1 | X | X | X | X | X | X |
| Medicago truncatula | 1 | X | X | X | X | X | X |
| Populus tremula | 1 | X | X | X | X | X | X |

| Table 2 | General information on the genomes of the studied organisms. |
|---------|-------------------------------------------------------------|
| Species | A% | C% | G% | T% | Chrom. no. | Genome size (bp) | No. of genes | Reference |
| Monocots | | | | | | | | |
| Brachypodium distachyon | 26.8 | 23.2 | 23.2 | 26.8 | 5 | 271,923,306 | 12,825 | [1] |
| Oryza sativa japonica | 28.2 | 21.8 | 21.8 | 28.2 | 12 | 382,150,945 | 54,814 | [2] |
| Oryza sativa indica | 26.6 | 21.4 | 21.4 | 26.6 | 19 | 427,026,737 | 49,710 | [3] |
| Z. mays | 26.5 | 23.5 | 23.5 | 26.5 | 10 | 2,065,722,704 | 54,814 | [4] |
| Triticum aestivum | 27.3 | 22.7 | 22.7 | 27.3 | 7 | 6,846,530,000 | n.a. | [5] |
| Dicots | | | | | | | | |
| Arabidopsis thaliana | 32.0 | 18.0 | 18.0 | 32.0 | 1 | 147,812,252 | 33,323 | [6] |
| Lotus japonica | 33.4 | 16.6 | 16.6 | 33.4 | 6 | 119,146,348 | 54,814 | [7] |
| Medicago truncatula | 33.4 | 16.6 | 16.6 | 33.4 | 9 | 307,511,856 | 49,710 | [8] |
| Populus tremula | 32.2 | 16.8 | 16.8 | 33.4 | 19 | 417,640,243 | n.a. | [9] |
occurred in the promoter region (3 kbp upstream region from the ATG start), within the gene body, and within the 3′ UTR (defined as the 3 kbp region downstream of the gene). The number of occurrences of each motif in each of the 3 gene subregions can be seen in Tables 5a and b. We measured how significantly different the number of occurrences was between the promoter region and the gene body and between the 3′ UTR and the gene body. We found that the difference was significant for promoters and 3′ UTRs in all six species. This shows that these top 15 motifs occur more significantly within promoters and 3′ UTR regions, both being regulatory regions, in these species.

**Motif content analysis of grass species**

Genetic similarity and therefore species relationships may be measured based on the similarity in the motif content. Species with higher content of common motifs with similar ranking according to motif score would be closer relatives than those species that have fewer common motifs. This is because the longer two species have been diverged from each other; more mutations have been allowed to accumulate between them, allowing a larger motif turnover to have taken place. The advantage of this method over sequence similarities between single genes is that it takes a global genomic sequence composition into account.

To test such phylogenetic sequence changes, we studied the number of common top 1000 genome motifs between the five monocot species. Two rice species belong to the family Ehrhartoideae, while Triticum and Brachypodium belong to the family Pooidae, and Zea belongs to the more distantly related family, Panicoideae. As we can see in Table 6, O. sativa and indica, the two rice species had the highest number of common genome motifs (939). The Spearman coefficient computed for these common motifs is also relatively high (0.710). The number of common motifs between the two Oryza species and Brachypodium is also proportionate (O. sativa japonica: 704 vs. O. sativa indica: 716). Brachypodium is also more related to the two rice species than it is to corn and wheat (only 414 and 498 common motifs, respectively). Conversely, Triticum has less motifs in common with the 2 rice species than does Brachypodium (446 and 448 for japonica and indica rice, respectively). This means that although Triticum and Brachypodium are in the same family, there might be a tradeoff when comparing their motif content with another species regarding the total number of motifs versus the motifs’ ranking. (See Table 5b.)

When comparing common motifs between the monocot species and an outlier species such as A. thaliana, we found that the number of common motifs was consistently low (Table 6). On average, there were 532 common motifs between any two monocot species, but only 422...
common motifs between A. thaliana and any given monocot species. In the case of the core promoter motifs (Fig. 2), we found only 10, 1, and 3 common motifs with Arabidopsis out of the top 100 with O. sativa japonica, indica, and B. distachyon. For the proximal promoters, this number was 11, 1, and 0. When comparing the number of common motifs between O. sativa japonica and A. thaliana within the 7 sequence sets, we also found them to be quite low (Fig. 3).

The scenario was somewhat different when the common top 1000 motif content of the core, proximal, and distal promoters were compared between the monocot species themselves (Tables 6a-c). The number of common motifs was greatest between the two rice species in the proximal and distal promoters (642 and 707, respectively). Surprisingly the number of common motifs in the core promoter between the two rice species was less than compared to either species versus Brachypodium. However, in the case of the core promoters and the 3’ UTRs (Table 7d), the Spearman correlation was greater when comparing the two rice species than when either rice species was compared to Brachypodium. About the same number of common motifs was found in all three species’ proximal promoters, and the Spearman coefficients were also roughly the same. This could be because the longer the promoter sequence gets, the larger the noise, allowing many more dissimilar statistically significant motifs to accumulate. (See Tables 7b and 7c.)

Analysis of promoter motifs in monocots

We compared the top 100 octamer motif content in the core, proximal, and distal promoters of the four monocot species that had promoter sequence sets available. The reason that this was not done for dicots was because only the genome sequence was available for those species. In the case of the core promoter motifs (Fig. 2), we found only 10, 1, and 3 common motifs with Arabidopsis out of the top 100 with O. sativa japonica, indica, and B. distachyon. For the proximal promoters, this number was 11, 1, and 0. When checking their annotations (the annotation of O. sativa japonica does not have good quality), we found 155 motifs from the top 100 with 1 species, 10 motifs with 2 species, 29 motifs with 3 species, 32 motifs with 4 species, and 19 motifs with 5 species. In the distal promoter sets, we found two pairs of paralog genes in O. sativa indica and B. distachyon, and 2 motifs in O. sativa japonica and T. aestivum had poor quality annotation; the wheat genome itself is very fragmentary and thus not usable for this analysis. We report a list of the top 50 genes from these 3 species in Supplementary Excel file 8 (SupplementaryFile8_MonocotPromoterMotifDistribution.xls) which have a high number of these motifs in their core, proximal, and distal promoters.

Here the Z. mays annotation proved to be fairly scant; however, we discovered a number of orthologous and paralogous gene sets in the proximal, and distal promoter gene sets. Orthologs are highlighted in bold in the Supplementary Excel file #8, whereas paralogs within species are underlined.

In the core promoter sets between O. sativa japonica and B. distachyon, we found a pair of transducin-WD-40 repeat family genes, 3 myb-like proteins, with 2 paralogs in Brachypodium. A RING/U-box superfamily gene pair was found in O. sativa japonica, and a pair of zinc-finger genes and a pentatricopeptide repeat gene pair was found in Brachypodium.

In the proximal promoter sets we found a WD-40 repeat gene in all 3 species, a pair of NB-ARC domain genes, RING/U-box genes, Class I peptidyl chain release factor genes, nodulin genes, purple acid phosphatase genes, and major facilitator superfamily genes in O. sativa japonica and B. distachyon. A pair of NAD[P+] oxidoreductase paralogs was found in O. sativa japonica, and a pair of glycosyltransferase genes was found in Brachypodium.

In the distal promoter sets, we found two pairs of paralog genes in O. sativa japonica; a pair of NB-ARC domain-containing disease resistance genes and a pair of cellulose synthase genes.
Table 5a

| Distribution of the 15 common motifs in three of the monocot species in promoters, within genes, and 3′ UTRs. |
|------------------------------------------------|
| promoter (p = 0.0017) | gene (p = 0.0017) | 3′ UTR (p = 0.0017) |
|----------------------|------------------|-----------------|
| Z. mays | B. distachyon | O. sativa |
| promoter (p = 6.7e-4) | gene | 3′ UTR |
|----------------------|------------------|-----------------|
| Z. mays | B. distachyon | O. sativa |
| AAAAGAAA | 2203 (35.44%) | 1715 (27.59%) | 2297 (36.95%) |
| AAAAGAAA | 2601 (35.28%) | 2068 (28.05%) | 2702 (36.65%) |
| AAAAGAAA | 2549 (35.57%) | 2014 (28.1%) | 2602 (36.31%) |
| AAAAGAAA | 2562 (36.46%) | 1911 (27.19%) | 2553 (36.33%) |
| AAAAGAAA | 2379 (35.15%) | 1922 (28.4%) | 2466 (36.44%) |
| AAAAGAAA | 2392 (35.38%) | 1935 (28.57%) | 2469 (36.44%) |
| AAAAGAAA | 2379 (35.15%) | 1922 (28.4%) | 2466 (36.44%) |
| AAAAGAAA | 2392 (35.38%) | 1935 (28.57%) | 2469 (36.44%) |
| AAAAGAAA | 2379 (35.15%) | 1922 (28.4%) | 2466 (36.44%) |
| AAAAGAAA | 2392 (35.38%) | 1935 (28.57%) | 2469 (36.44%) |
| AAAAGAAA | 2379 (35.15%) | 1922 (28.4%) | 2466 (36.44%) |
| AAAAGAAA | 2392 (35.38%) | 1935 (28.57%) | 2469 (36.44%) |
| AAAAGAAA | 2379 (35.15%) | 1922 (28.4%) | 2466 (36.44%) |
| AAAAGAAA | 2392 (35.38%) | 1935 (28.57%) | 2469 (36.44%) |
| AAAAGAAA | 2379 (35.15%) | 1922 (28.4%) | 2466 (36.44%) |
| AAAAGAAA | 2392 (35.38%) | 1935 (28.57%) | 2469 (36.44%) |

Motif pair analysis

The distribution of motif pairs within individual sequences belonging to the core, proximal, and distal promoters, 5′ UTRs, 3′ UTRs, and 1st introns was also examined in the monocot species for which these sequence sets were available. The top 5050 possible motif pairs were examined, coming from all possible pairing of the top 100 motifs from each sequence set from each species. The sequences for all of these motif pairs, their real and expected occurrences and their motif pair score can be found in the supplementary Excel files for each monocot species (Supplementary files 1-5). The reason dicot genomes were not examined is because motif pairs form the building blocks of transcription factor modules, and as such, require being in close proximity with each other, and this is not possible with the genome where motifs would be far away because of the sheer size of the genome.

The number of common motif pairs was examined for the core, proximal, and distal promoters, and 3′ UTRs, as can be seen in Supplementary Tables 7a-d. The Spearman correlation coefficient was also calculated for the common motif pair content. As with the Spearman correlation coefficient calculations performed with only single motifs, a trade-off between the number of common motif pairs and the Spearman correlation between them. It occurs many times that there are more motif pairs between either rice species and B. distachyon or Z. mays, compared to the number of common motifs between the two rice species; however, the Spearman coefficient is still larger between the two rice species. Conversely, the opposite is true, that there are less motif pairs between either rice species and B. distachyon and Z. mays; however, the Spearman coefficient is greater between these species and both rice species. This may be due to the fact that it is more likely to get a higher Spearman coefficient value with fewer common elements than with a larger number of common elements, which might easily introduce noise.

Otherwise, the similarity between the two rice species and Z. mays can be seen in that the common motif pair content drops off sharply to only 231 (4.6% of all 5,050 pairs) between O. sativa indica, and 351 (7% of all 5050 pairs) between B. distachyon and maize. The number is also relatively low between these two previously mentioned species when looking at the common motif pairs in proximal promoters (820, which is 16.2% of all 5050 pairs), and in distal promoters (990, which is 19.6% of all 5050 pairs).

Dicot consensus sequences

For the dicot species we defined consensus sequences for the top 100 genome. This we did by putting motifs with similar sequences into the same cluster. Afterwards, a multiple alignment was made for each cluster, and the consensus sequence was defined from the alignment. These consensus sequences can be seen in the tab “consensus sequences” in Supplementary File #6.
Conclusion

Comparing the motif content of different species at different taxonomical levels broadens our horizons and allows a deeper analysis, making it possible to uncover newer aspects of their genomes. For example, by looking at the motif content of whole genomes, we can capture global similarity between them in a holistic manner, instead of looking at similarities between genes on only a local level. This is especially evident in the number of common top-ranking motifs, for example within the studied monocot species, which reflect the degree of relationships between individual species. Using the top-ranking motifs from the whole genome is especially useful for this, compared to the motif content of different parts of the gene (such as the promoter or 3' UTR), since this way we can draw general conclusions about the whole genome.

The similar content of top-ranking motifs between two given species can be quantified to measure how similar the two species are to one another. This can be done by comparing the number of common motifs, and/or by calculating the Spearman coefficient between sets of motifs between the two species. The number of common motifs between two sets shows how similar two species are, because we would expect that the closer two species are to one another, they would share more motifs in common which also occur more abundantly, and thus have a high score. Using the Spearman coefficient also gives a good measure of this similarity. This can be seen in the case of motif content similarity in the 3' UTR regions between the two rice species compared to *B. distachyon*. Here even though there are a smaller number of common motifs between the two rice species compared to *B. distachyon* (167 compared to 376 and 515), their Spearman coefficient is larger (0.947 compared to 0.589 and 0.498). This also might be due to less selective pressure on the 3' UTR region allowing for larger sequence divergence. Taking both measures into account shows how related two species are to each other. When comparing the common motif content with an outlier species such as *A. thaliana*, which consistently gave a low number of common motifs as well as a low Spearman coefficient.

For further study, it would be intriguing to analyze the motif content of a large number of species (30 or more), in order to perform motif content analysis on a larger scale. To this end we have written a perl script (genomotifome.pl) which analyzes the whole genome sequence of a given species for motifs of a given length, returning a set of the top-ranking motifs from that genome.

Furthermore, the analysis of top-ranking motifs also makes it possible to predict parts of secondary structures of different kinds of genetic elements, such as miRNA sequences, tandem repeats, or microsatellites. This is a novel property of the analysis of the motif content of different parts of the genome, and can complement other algorithms or models which do the same. We were also able to find sets of motifs which have a significantly higher occurrence in regulatory regions (promoters and 3' UTRs). Furthermore, by analyzing the distribution of conserved motifs common to different species, we were able to find genes which could possibly take part in the same physiological processes under the same regulation.

In summary, motif cross-comparison between a number of different plant species provides new and exciting results which can be applied and broadened to other organisms as well to deepen our understanding of the regulation of their genomes.

Materials and methods

Sequence sets

Links and references to the data sets used in this analysis can be seen in Table 1. Here general information on the genomes of each species is listed, such as the number of chromosomes, number of genes, and total genome size. For the dicot species excluding *Arabidopsis* and also for the monocot, *T. aestivum*, only the whole genome sequence was available (for *Triticum* it was available only in contig sequences). *B. distachyon* had only the whole genome, the core promoter and proximal promoter sets and the 3' UTRs available, while *Z. mays* had only the whole genome sequence, and the core, proximal and distal promoter sets available. While these sequence sets may be incomplete as of yet, we felt it worthwhile to analyze the available data and try to draw conclusions from them through multiple species comparisons.

The genome sequence of *Arabidopsis* has already been done by Lichtenberg at al. [9], however we performed our own analysis on the genome of this plant species, thus we were working with our own *Arabidopsis* data sets. Since the analysis had already been performed for *O. sativa japonica* by Cserhati [8], we simply used the data sets already available for that species. The genomes for eight of the nine species were masked using RepeatMasker [24] in order to purge them of repeats. There were technical difficulties with the *T. aestivum* genome due to the fragmented nature of the genome as it had not yet been assembled into whole chromosome sequences, the program took too long to run.

| Motif | Arabidopsis promoter (p = 5.1e-4) | Arabidopsis gene | Arabidopsis 3' UTR (p = 5.9e-4) | Medicago promoter (1.9e-3) | Medicago gene | Medicago 3' UTR (2e-3) |
|-------|---------------------------------|-----------------|---------------------------------|--------------------------|-------------|----------------------|
| AAAAAA | 113367 (49.58%) | 21263 (9.3%) | 93997 (41.11%) | 176179 (37.06%) | 123956 (26.08%) | 175149 (36.85%) |
| AAAAAGA | 20023 (44.96%) | 6323 (14.19%) | 18188 (40.84%) | 21275 (36.99%) | 15009 (26.09%) | 21224 (36.9%) |
| AAAAAGAA | 23235 (46.66%) | 7359 (14.14%) | 21421 (41.18%) | 25438 (36.84%) | 18090 (26.19%) | 25518 (36.95%) |
| AAAAATAA | 27470 (50.34%) | 5025 (9.25%) | 22073 (40.45%) | 46731 (37%) | 32212 (26.29%) | 46353 (36.7%) |
| AAACAAA | 24919 (44.73%) | 7887 (14.15%) | 22895 (41.1%) | 27648 (37.09%) | 19527 (26.19%) | 27365 (36.71%) |
| AAATAAA | 26280 (49.71%) | 4922 (9.31%) | 21657 (40.97%) | 43508 (36.98%) | 31295 (26.35%) | 43526 (36.65%) |
| AAAAAAG | 25016 (45.3%) | 7900 (14.3%) | 23202 (40.38%) | 28103 (40.92%) | 20056 (26.35%) | 27943 (36.71%) |
| TICTITTTT | 19697 (49.91%) | 6015 (13.71%) | 18147 (41.37%) | 21078 (37.14%) | 14781 (26.04%) | 20885 (36.8%) |
| TITTTTTT | 26774 (49.75%) | 5057 (9.39%) | 21986 (40.85%) | 46196 (36.8%) | 33009 (26.29%) | 46310 (36.89%) |
| TTCTTTTT | 23991 (44.74%) | 7315 (14.17%) | 21204 (41.08%) | 25449 (36.95%) | 17949 (26.06%) | 25458 (36.97%) |
| TTGTGTTT | 26116 (45.24%) | 8194 (14.2%) | 23419 (40.57%) | 24684 (36.71%) | 17745 (26.39%) | 24814 (36.9%) |
| TTGTITTT | 25977 (49.31%) | 4984 (9.46%) | 21719 (41.22%) | 43667 (36.87%) | 30965 (26.14%) | 43790 (36.79%) |
| TTGTCTTT | 24914 (45.01%) | 7814 (14.11%) | 22619 (40.86%) | 27275 (36.84%) | 19404 (26.2%) | 27354 (36.94%) |
| TTGTCTTTT | 25058 (45.5%) | 7743 (14.05%) | 22271 (40.43%) | 28005 (36.85%) | 19829 (26.09%) | 28143 (37.04%) |
| TTGTCTTTTT | 112239 (49.57%) | 20617 (9.18%) | 93565 (41.32%) | 173501 (36.76%) | 123670 (26.2%) | 174684 (37.02%) |
Motif statistical measure

The statistical significance of a given motif \( m \) is \( \text{sign}(m) = S \cdot \ln(S/ES) \), where \( S \) is the number of sequences the motif \( m \) occurs in, and \( ES \) is the number of sequences the motif is expected to occur in, by calculating the probability of the motif’s occurrence based on the background base distribution in a given species. The probability \( p_m \) can be calculated with the following formula: \( p_m = \prod_{i=1}^{n} p_{X,i} \), where \( n \) is the length of the motif, \( i \) is a running variable from 1 to \( n \), and \( p_{X,i} \) is the \( i \)th base in the motif, where \( X = \{A,C,G,T\} \), and \( n \) is the length of the motif (6–10 for hexamers to decamers). Only motifs not containing ambiguous IUPAC letters (M, R, W, S, Y, K, B, D, H, V, or N) were retained. Thus, motifs which are overrepresented in the genome (compared to their expected occurrence) will receive a higher score. Not only are those motifs scored higher which occur relatively more than their expected value (\( S/ES \)), but especially those with a high occurrence in general (that is why we multiple \( \ln(S/ES) \) with \( S \), the number of occurrences).

In the whole genome, the expected occurrence of \( w \) is \( ES(m) = N_{\text{genome}} \times p_m \), where \( N_{\text{genome}} \) is the size of the species’ genome, and \( p_m \) is the occurrence probability of the motif. In the case of the other

Fig. 2. Pairwise comparison of common putative motifs in the whole genome, core and proximal promoters between all monocot and dicot species.

Fig. 3. Number of common motifs within the 7 sequence subsets between Oryza sativa japonica and Arabidopsis thaliana.
Table 7a
Core promoter motif content similarity and Spearman ranking between the studied monocot species.

|                | O. sativa japonica | O. sativa indica | B. distachyon | Zea mays |
|----------------|-------------------|-----------------|---------------|----------|
| O. sativa japonica | 271 (0.688)      | 404 (0.531)     | 432 (0.544)   |          |
| O. sativa indica  | 515 (0.498)      | 569 (0.420)     |               |          |
| B. distachyon     | 580 (0.479)      |                 |               |          |
| Zea mays          |                   |                 |               |          |

Table 7b
Proximal promoter motif content similarity and Spearman ranking between the studied monocot species.

|                | O. sativa japonica | O. sativa indica | B. distachyon | Zea mays |
|----------------|-------------------|-----------------|---------------|----------|
| O. sativa japonica | 642 (0.344)      | 645 (0.416)     | 655 (0.409)   |          |
| O. sativa indica  | 686 (0.399)      | 439 (0.564)     |               |          |
| B. distachyon     | 518 (0.447)      |                 |               |          |
| Zea mays          |                   |                 |               |          |

Table 7c
Distal promoter motif content similarity and Spearman ranking between the studied monocot species.

|                | O. sativa japonica | O. sativa indica | Zea mays |
|----------------|-------------------|-----------------|----------|
| O. sativa japonica | 707 (0.341)      | 619 (0.384)     |          |
| O. sativa indica  | 460 (0.545)      |                 |          |
| Zea mays          |                   |                 |          |

Determination of consensus sequences in dicots

For the four dicot species, the top 100 octamers were all compared with one another to determine which sequence was similar to the other, and to what degree. Clusters of motifs from the top 100 motifs were defined where the members of the cluster were similar to at least one other cluster member by a sequence similarity of at least 87.5% (allowing for 1 mismatch). For each cluster the motif members were put into a multi-fasta file and the ClustalW2 program was run on it to determine a consensus sequence for that motif.

Other data and methods

Gene annotations for O. sativa japonica and B. distachyon were downloaded from http://www.plantgdb.org. A list of miRNA sequences and their targets can be found at http://bioinformatics.cau.edu.cn/PMRD/adjunct/osa_miR_target.txt. A multifasta sequence file containing known microRNA sequences was downloaded from http://bioinformatics.cau.edu.cn/PMRD/adjunct/osa_mature [23]. Gtf files for O. sativa japonica and indica, T. aestivum, Z. mays, B. distachyon, A. thaliana, and M. truncatula were all downloaded from http://plants.ensembl.org/info/website/ftp/index.html. The conversion between O. sativa japonica gene ids was performed with this file: http://rapd. dna.aﬀrc.go.jp/download/archive/RAP-MSU.txt.gz. p-Values for the differences in motif numbers between promoters and 3’ UTRs versus gene were calculated by the Wilcoxon-test in R.

Venn diagrams

The Venn diagrams were calculated using the software at the Bioinformatics and Evolutionary Genomics Workgroup at http://bioinformatics.psb.ugent.be/cgi-bin/liste/Venn/calculate_venn.html.

PLACE motif definitions

The supplementary dictionary for defining the functions of the PLACE database motif ids was adapted from http://ftp.dna.aﬀrc.go.jp/pub/dna_place/place.fasta.

Perl script

The perl script can be downloaded from the author’s webpage at http://unmc.edu/bsb/c/docs/motifome_script.zip.
### References

1. Web Reference 1. [http://www.brachypodium.org/](http://www.brachypodium.org/).
2. Web Reference 2. [http://rice.genomics.org.cn/rice/index2.jsp](http://rice.genomics.org.cn/rice/index2.jsp).
3. Web Reference 3. [ftp://ftp.plantdb.org/download/Genomes/ZmGDB/](ftp://ftp.plantdb.org/download/Genomes/ZmGDB/).
4. Web Reference 4. [http://www.cerealsdb.uk.net/CerealsDB/Documents/D0C_copy-right.php](http://www.cerealsdb.uk.net/CerealsDB/Documents/D0C_copy-right.php).
5. Web Reference 5. [ftp://ftp.plantdb.org/download/Genomes/PtGDB/](ftp://ftp.plantdb.org/download/Genomes/PtGDB/).
6. Web Reference 6. [http://www dna.affrc.go.jp/PLACE/](http://www dna.affrc.go.jp/PLACE/).

### Table 8

| Gene ID          | Number of reverse complementary 3′ UTR motifs | Gene annotation                                                                 |
|------------------|-----------------------------------------------|--------------------------------------------------------------------------------|
| Bradi1g2590      | 53                                            | 6-Phosphogluconate dehydrogenase family protein                                 |
| Bradi1g56250     | 60                                            | A20/AN1-like zinc finger family protein                                          |
| Bradi4g15400     | 73                                            | Agerat domain-containing protein                                                 |
| Os01g18729       | 75                                            | ARM repeat superfamily protein                                                   |
| Os08g43090       | 64                                            | Basic-leucine zipper (bZIP) transcription factor protein                         |
| Bradi1g11310     | 61                                            | B-box-type zinc finger protein with CCT domain                                   |
| Os12g16350       | 57                                            | Beta-hydroxyisobutyr-CoA hydrase 1                                              |
| Bradi3g30120     | 61                                            | Beta-ketocarboxylate 2                                                          |
| Os01g27530       | 74                                            | Calcium-binding EF-hand family protein                                           |
| Bradipg24460     | 61                                            | Calmodulin-binding protein                                                      |
| Os08g03100       | 67                                            | CCCH-type zinc finger family protein with RNA-binding domain                     |
| Os02g57410       | 75                                            | Cysteine proteinases superfamily protein                                         |
| Bradi3g01980     | 81                                            | Cysteine proteinases superfamily protein                                        |
| Os01g10040       | 55                                            | Cytochrome P450, family 90, subfamily D, polypeptide 1                           |
| Os11g09570       | 72                                            | FAD/NAD(P)-binding oxidoreductase family protein                                 |
| Os01g08830       | 54                                            | F-box family protein with a domain of unknown function (DFU295)                 |
| Bradi2g88800     | 61                                            | F-box/RNI-like superfamily protein                                              |
| Bradipg05226     | 64                                            | Gigantea protein (GI)                                                           |
| Os11g47870       | 50                                            | GRAS family transcription factor                                                 |
| Os08g33750       | 58                                            | Homedomain-like superfamily protein                                             |
| Os06g06809       | 59                                            | Hydroxylase-like family protein                                                 |
| Os02g01150       | 61                                            | Hydroxysteroid reductase                                                        |
| Os04g56500       | 52                                            | IL11 binding BH1 (I)                                                            |
| Os01g61720       | 54                                            | IQ-domain 2                                                                     |
| Os01g10504       | 87                                            | K-box region and MADS-box transcription family protein                          |
| Os07g01490       | 156                                           | Kinesin 5                                                                       |
| Os10g13970       | 68                                            | Leucine-rich repeat protein kinase family protein                                 |
| Os06g19990       | 56                                            | LORELEI-LIKE-GLP ANCHORED PROTEIN 3                                             |
| Os08g45380       | 52                                            | LRR and NB-ARC domains-containing disease resistance family protein             |
| Os07g44090       | 60                                            | mbp domain protein 61                                                           |
| Os01g09550       | 65                                            | NAC domain containing protein 75                                                 |
| Bradi1g08890     | 51                                            | PEBP (phosphatidylyethanolamine-binding) protein family                         |
| Bradi1g04820     | 56                                            | Peptidease S24/S26A/S26B/S26C family protein                                     |
| Os01g33880       | 60                                            | Phytochrome-associated protein 1                                                |
| Os12g37480       | 53                                            | Plant inverse/pectin myristeerase inhibitor superfamily protein                  |
| Os02g11000       | 59                                            | Plant Tudor-like RNA-binding protein                                             |
| Os06g41930       | 50                                            | PLATZ transcription factor family protein                                         |
| Os07g28260       | 72                                            | P-loop containing nucleoside triphosphate hydrolases superfamily protein         |
| Os05g01380       | 53                                            | Polyglutamatecontaining protein inhibited 1                                     |
| Os03g57940       | 56                                            | Protein kinase family protein                                                    |
| Os04g21340       | 52                                            | Protein of unknown function (DFU1685)                                           |
| Os08g45170       | 74                                            | Protein of Unknown Function (DFU239)                                            |
| Os02g08364       | 76                                            | Protein phosphatase 2C family protein                                            |
| Bradi2g30860     | 67                                            | Protein with RING-U-box and TRAF-like domains                                   |
| Bradi3g52740     | 57                                            | Pyrophosphorylase 1                                                             |
| Bradi2g40400     | 63                                            | Ribosomal L28 family                                                            |
| Os06g03580       | 58                                            | RING-U-box superfamily protein                                                  |
| Bradi4g45400     | 56                                            | Saposin B domain-containing protein                                             |
| Os03g27590       | 58                                            | Serine carboxypeptidase-like 51                                                |
| Bradi2g39275     | 152                                           | Serine protease inhibitor, potato inhibitor I-type family protein                |
| Bradi1g21510     | 80                                            | SPX domain gene 3                                                               |
| Bradi5g01823     | 61                                            | Terpenoid cyclases/protein                                                      |
| Os02g36210       | 57                                            | Terpenoid transferases superfamily protein                                       |
| Os12g08260       | 63                                            | Thiamin diphosphate-binding fold (THDP-binding) superfamily protein             |
| Os04g20400       | 118                                           | UDP-Glycosyltransferase superfamily protein                                      |

### Appendix A. Supplementary data

Supplementary data is available for all the calculations done on *Oryza sativa indica*, *Brachypodium distachyon*, *Triticum aestivum*, *Zea mays*, *Arabidopsis thaliana*, *Lotus japonicus*, *Medicago truncatula*, and *Populus tremula*. For supplementary information on *Oryza sativa indica*, the reader is referred to [8]. Supplementary data to this article can be found online at [http://dx.doi.org/10.1016/j.gdata.2014.12.006](http://dx.doi.org/10.1016/j.gdata.2014.12.006).