Published online 12 September 2007 *Nucleic Acids Research, 2007, Vol. 35, No. 18* ***6219–6226***
doi:10.1093/nar/gkm685

Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis

Yoshiharu Y. Yamamoto[1,*], Hiroyuki Ichida[2,3], Tomoko Abe[2], Yutaka Suzuki[4], Sumio Sugano[4] and Junichi Obokata[1]

[1]Center for Gene Research, Nagoya University, Nagoya, Aichi 464-8602, [2]RIKEN Nishina Center for Accelerator-Based Science, 2-1 Hirosawa, Wako, Saitama 351-0198, [3]Graduate School of Science and Technology, Chiba University, 1-33, Yayoi-cho, Inage-ku, Chiba-shi, Chiba 263-8522 and [4]Department of Medical Genome Sciences, Graduate School of Frontier Science, Institute of medical Science, University of Tokyo, Shiroganedai 4-6-1, Minato-ku, Tokyo 108-8639, Japan

Received August 1, 2007; Revised August 19, 2007; Accepted August 20, 2007

ABSTRACT

Mammalian promoters are categorized into TATA and CpG-related groups, and they have complementary roles associated with differentiated transcriptional characteristics. While the TATA box is also found in plant promoters, it is not known if CpG-type promoters exist in plants. Plant promoters contain Y Patches (pyrimidine patches) in the core promoter region, and the ubiquity of these beyond higher plants is not understood as well. Sets of promoter sequences were utilized for the analysis of local distribution of short sequences (LDSS), and approximately one thousand octamer sequences have been identified as promoter constituents from *Arabidopsis*, rice, human and mouse, respectively. Based on their localization profiles, the identified octamer sequences were classified into several major groups, REG (Regulatory Element Group), TATA box, Inr (Initiator), Kozak, CpG and Y Patch. Comparison of the four species has revealed three categories: (i) shared groups found in both plants and mammals (TATA box), (ii) common groups found in both kingdoms but the utilized sequence is differentiated (REG, Inr and Kozak) and (iii) specific groups found in either plants or mammals (CpG and Y Patch). Our comparative LDSS analysis has identified conservation and differentiation of promoter architectures between higher plants and mammals.

INTRODUCTION

Transcription of structural genes is directed by the corresponding promoters, and their DNA sequence encodes timing, strength, direction and position of transcriptional initiation. Therefore, decoding the promoter sequence is one of the most important issues in genome biology.

A typical promoter is composed of a core promoter and regulatory domains. Several core promoter elements have been identified: the TATA box, the TFIIB-Recognition Element (BRE), Inr, the Downstream Promoter Element (DPE) for *Drosophila* promoters and CpG islands (1). They function cooperatively or opposingly depending on their characteristics (1,2).

Studies on comprehensive mapping of transcription start sites (TSSs) of mammals have revealed that a promoter often contains multiple TSSs as a cluster, and the shape of the cluster, as well as the profile and strength of expression is reflected by the type of the core promoter. Mammalian TATA-type promoters show tissue-specific expression and sharp TSS cluster with one major TSS, while CpG-associated promoters, found more in TATA-less promoters, tend to show ubiquitous rather than regulated gene expression and broad-type TSS clusters (3,4). These observations suggest differentiated and complementary roles between the TATA box and the CpG island in mammals. Recently, differential mutation rates of promoters according to the core types have been observed (5), indicating that the core type of a promoter determines not only transcriptional characteristics but also evolutional tendency.

In spite of the various identified core promoter elements mentioned above, it is not known if all the promoters in a genome can be explained by these core elements, or whether unknown elements are still to be found. In addition, much less information is available for plant core promoters.

Regulatory regions are often found upstream of the core promoter region. A systematic deletion analysis of 387 human promoters by Cooper *et al.* suggests that a 1 kb promoter can be divided into two regulatory regions: −1000 to −500 bp as a negative regulatory region and −300 to −50 bp for positive contribution to core promoter

*To whom correspondence should be addressed. Tel: +81 52 789 3083; Fax: +81 52 789 3083; Email: yyoshi@gene.nagoya-u.ac.jp

activity (6). Another report on the distribution of human regulatory elements suggests that they make clusters at the 5′ regions of genes as well as at the 3′ regions (7). In addition, so-called long-range enhancers which act from as far away as 1 Mb are also reported (8,9).

The structure of a promoter is recognized by the presence of known promoter elements. Therefore, accurate recognition of a promoter structure relies on a comprehensive list of promoter elements. The availability of complete genome sequence has led to the development of approaches utilizing bioinformatics to extract promoter elements: extraction of consensus sequences from a set of co-regulated promoters [Gibbs Motif Sampling (10,11), MEME (12)], searches for over-represented sequences in a co-regulated promoter set over a reference set (13–15), and identification of conserved sequences at the promoter region by comparative genomic approaches (16–19). The discovery of preferential localization of several transcription factor-binding sequences along human promoters (20) opened up a novel approach to extract elements according to distribution profiles along promoters (21,22). The advantages of this method are: (i) sensitive detection of minor elements in the analyzed promoter set, which is superior to the extraction of consensus sequences, (ii) no knowledge of co-regulated gene sets is required and (iii) the availability of single genome-specific information. A combinational approach of detection of consensus sequences and consideration of their positional information is also reported [A-GLAM, (23)].

Promoter elements identified by any of these approaches can be used for promoter annotation, classification and prediction (24,25). Compared with the rich information available on animal promoters, plant promoter architecture is poorly understood, and in particular, information about the core region is sparse. In order to understand plant promoter architecture, we applied the extraction method with the distribution profile to *Arabidopsis* and rice promoters [LDSS (local distribution of short sequences) analysis, (26) to detect REG (Regulatory Element Group), TATA and Y Patch (pyrimidine patch)] groups. The preferred sequences of *Arabidopsis* and rice were moderately differentiated, but essentially all the major groups were conserved. It was expected that the REG and TATA groups would be found in plant promoters as well as in animal promoters, but there were no reports of Y Patch from an animal genome. Therefore, we decided to investigate if it can also be detected in mammalian promoters by the same approach. In addition, we addressed if CpG-type promoters also exist in plants as well as in mammals. Our comparative analyses have revealed conservation and differentiation of the promoter architecture between plants and mammals. This is the first report on the differentiation of core promoter architectures between higher plants and mammals.

MATERIALS AND METHODS

Sequence analysis

Preparation of *Arabidopsis* and rice promoter sequences are described elsewhere (26,27). A total of 30 957 human promoters sequences and 18 088 mouse sequences were obtained from DBTSS (28) (http://dbtss.hgc.jp). All the utilized promoter sequences are based on experimentally identified TSSs.

Extraction of LDSS-positive octamer sequences was done as described elsewhere (26). Clustering of the octamers according to distribution profiles was achieved using the Cluster software (http://rana.lbl.gov/EisenSoftware.htm) with the *k*-means method, and visualized by TreeView (http://rana.lbl.gov/EisenSoftware.htm). Clustering was achieved by two sequential steps, where related clustering groups were joined and then subjected to a second clustering analysis. The other sequence analyses were achieved with the aid of perl and C++ programs and also Excel software (Microsoft Japan, Tokyo).

Mammalian regulatory motifs were identified with the following information: CCAAT motif (CCAAT) (29), CRE-related motif [TGACGT, cAMP responsive element (21)], ETS [G/CCGGAA, regulation of embryonic and adult hematopoiesis (21,30)], NRF-1 [CGCATG, regulation of mitochondrion targeting protein genes (31)].

RESULTS AND DISCUSSION

Extraction and grouping of LDSS-positive octamers from plants and mammals

In our previous report, *Arabidopsis* and rice promoters from −1000 to −1 region were analyzed. This work extends the analyzed region from −1 to +200. All the possible octamer sequences (4^8 = 65 536) were examined for distribution profiles along the −1000 to +200 region relative to identified TSSs, and ones with localized distribution were extracted according to the method called LDSS analysis as described previously (26). 1719, 1148, 1597 and 2719 sequences were identified as LDSS-positive octamers from *Arabidopsis*, rice, human and mouse, respectively. These octamers were subsequently subjected to clustering analysis according to their distribution profiles.

Recently, Kiran *et al*. reported the effects of extensive point mutations of a typical TATA element on transcriptional activity measured by tobacco transient assays (32). We found that their results are nicely explained by our LDSS analysis for *Arabidopsis* TATA sequences (Table S1). These results clearly indicate the functional implication of LDSS parameters.

Characterization of the clustered groups

Clustering analysis has revealed several octamer groups that share distribution profiles (Figure 1). As reported previously (26), REG (Regulatory Element Group), TATA and Y Patch (pyrimidine patch) groups are found in *Arabidopsis* and rice. The TATAD group (Figure 1, and Table S2 for the sequence) appears between the TATA box and the TSS. In this report, we extended the analyzed promoter region from −1 to +200. This extension resulted in the identification of Initiator (Inr) that is a consensus around the TSS, and a consensus for translational initiation (Kozak). In addition, minor motifs shown as CA, GA, CAAC and Down are also recognized in this study (Figure 1, and Table S2 for sequence).

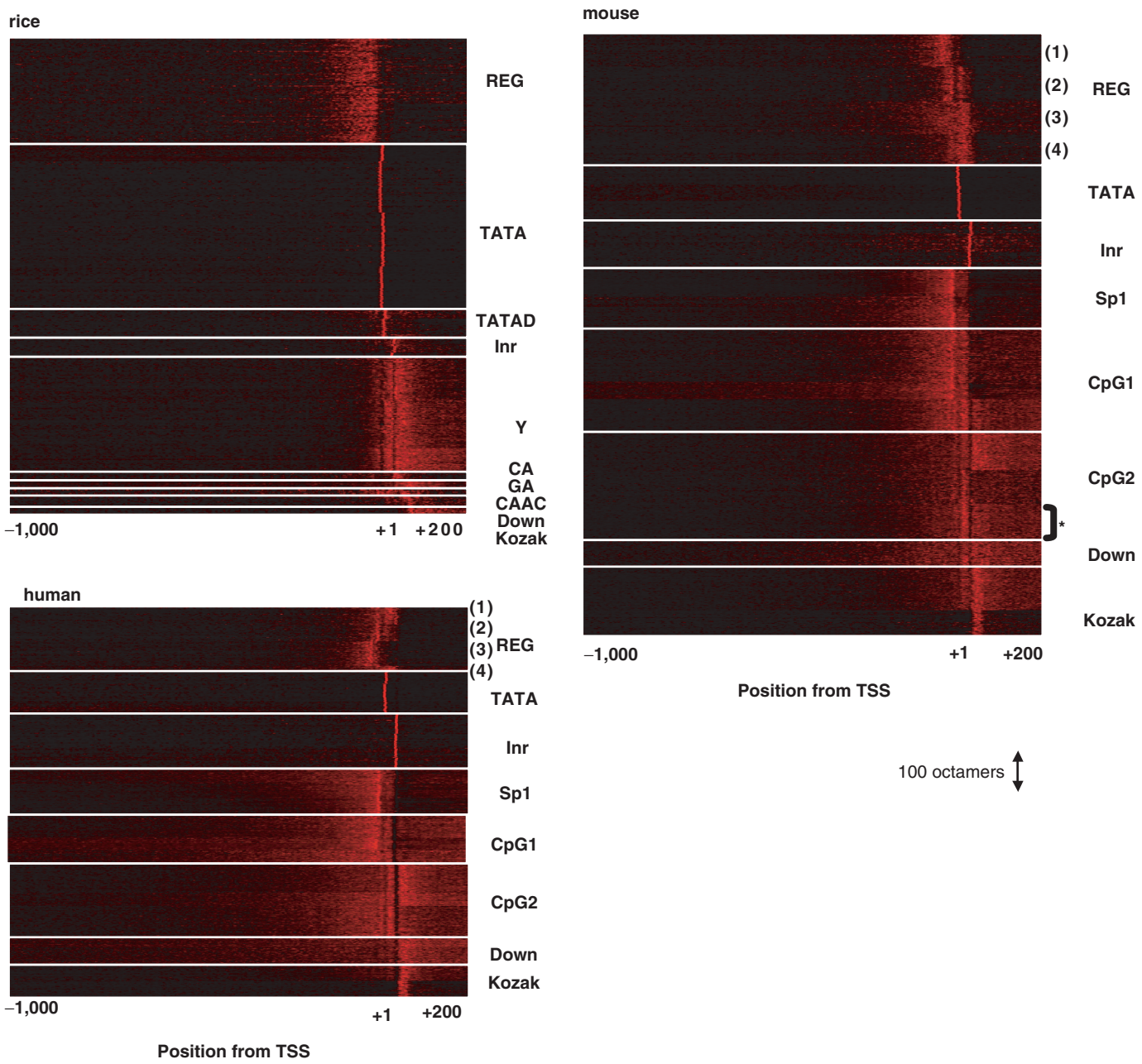

Figure 1. LDSS-positive octamer groups. Distribution profiles of the clustered LDSS-positive octamers from *Arabidopsis*, rice, human and mouse are shown. The arrow in the figure indicates the width for 100 ocatmers.

In order to focus on promoter architecture in relation to transcriptional regulation, groups with distribution peaks downstream of Kozak are neglected in this study. According to the extension of the analyzed region from our previous studies on rice (26) and subsequent clustering, TATAD, Inr, CA, GA, CAAC, Down and Kozak groups have been newly recognized in this study, and new members of Y Patch have been detected.

As expected, human and mouse share REG, TATA, Inr and Kozak with plants (Figure 1). However, mammalian groups lack the Y Patch. Instead, CG-rich group (CpG1 and CpG2, Table S3 for sequence), that have similar distribution profiles to plant Y Patch groups, are detected (Figure 1). Sequences of the detected groups by LDSS analysis are shown in Table S4 (human) and Table S5 (mouse), and summarized in Table 1. The summary table also shows directional sensitivity of the groups. REG and Sp1 are both direction insensitive, which means the complementary sequence of a member also shows the same distribution profile. TATA, Y Patch, Inr and Kozak are direction sensitive. As shown in the table, the directional characteristics of each group are conserved among plant and mammalian species. It should be mentioned that the identified direction sensitivity does not conflict with our biological knowledge on the indicated elements.

Table 1. Detection of major groups by LDSS analysis

	Arabidopsis	Rice	Human	Mouse	% comp[a]	Direction sensitivity[b]
REG	Yes	Yes	Yes	Yes	41–57	No
TATA	Yes	Yes	Yes	Yes	6–12	Yes
Sp1	No	No	Yes	Yes	56–64	No
CpG	No	No	Yes	Yes	35–37	Yes/no
Y Patch	Yes	Yes	No	No	0	Yes
Inr	Yes	Yes	Yes	Yes	0–11	Yes
Kozak	Yes	Yes	Yes	Yes	0–13	Yes

[a]Percentage of sequences whose complementary sequences are also found in the group.
[b]Judged as no sensitivity if % comp is over 50.

Taking the groups found in both plants and mammals, we carried out analysis to determine how many octamer sequences are conserved (Figure 2). As shown in Panel A, sequence conservation is most obvious for the TATA groups in both plants and mammals, and the Inr and REG groups are less conserved, except for mammalian REG groups. Conservation between plants and mammals is shown in Panel B, and the tendency observed in Panel A is again obvious by comparison of the plant and mammalian groups. These results indicate that TATA sequences are well conserved between plants and mammals, while Inr and REG are not. Although there is no shared octamer for plant and mammalian Inr, a dimer consensus at the $-1/+1$ position, the YR Rule, where Y (C or T) at the -1 position and R (A or G) at $+1$, is applicable to both mammals (4) and plants (26). Therefore, the consensus for TSSs at a dimer level is essentially conserved between plants and mammals.

In summary, common promoter element groups have been revealed from plant and mammalian promoters, judged by shared distribution profiles and direction sensitivities. The most conserved group is the TATA box whose sequences are also well conserved. The other common groups, REG and Inr, have differentiated sequences between plants and mammals.

Promoter elements including regulatory and core ones are recognized by *trans*-factors, that are DNA-binding proteins in most cases. Therefore, conservation of TATA box between higher plants and mammals would be a reflection of conservation of the TATA-binding proteins, and divergence of REG would be due to diversification of DNA-binding transcription factors. This idea is supported by the fact that TATA-binding proteins have been found from *Arabidopsis*, human and also yeast with high conservation (33), and that ~45% of the *Arabidopsis* transcription factors belong to plant-specific gene families (34).

Subgroups of mammalian REG

We noticed that mammalian REGs could be further classified into several subtypes according to distribution profiles (Figure 1, REG). The mouse octamer sequences of these subgroups are shown in Table S6. As seen in the table, most octamers of mouse REG1 are CCAAT boxes, whereas in REG2 CRE is common but CCAAT is very rare, indicating a certain relationship between distribution profiles and sequence motifs. After trails of classification by *k*-means clustering with several *k* values, we found that clustering into four groups resulted in the most uneven or distorted separation of motifs for human and mouse REGs. The distributional differentiation of mouse *cis*-elements is summarized in Figure 3. Our data indicates that a CCAAT-binding factor comes upstream of ETS if they co-exist in a promoter (Figure 3A). This differentiation of the mammalian REG may facilitate functional cooperation of different types of transcription factors. This phenomenon appears to be specific to mammals and is not observed in the plant REG.

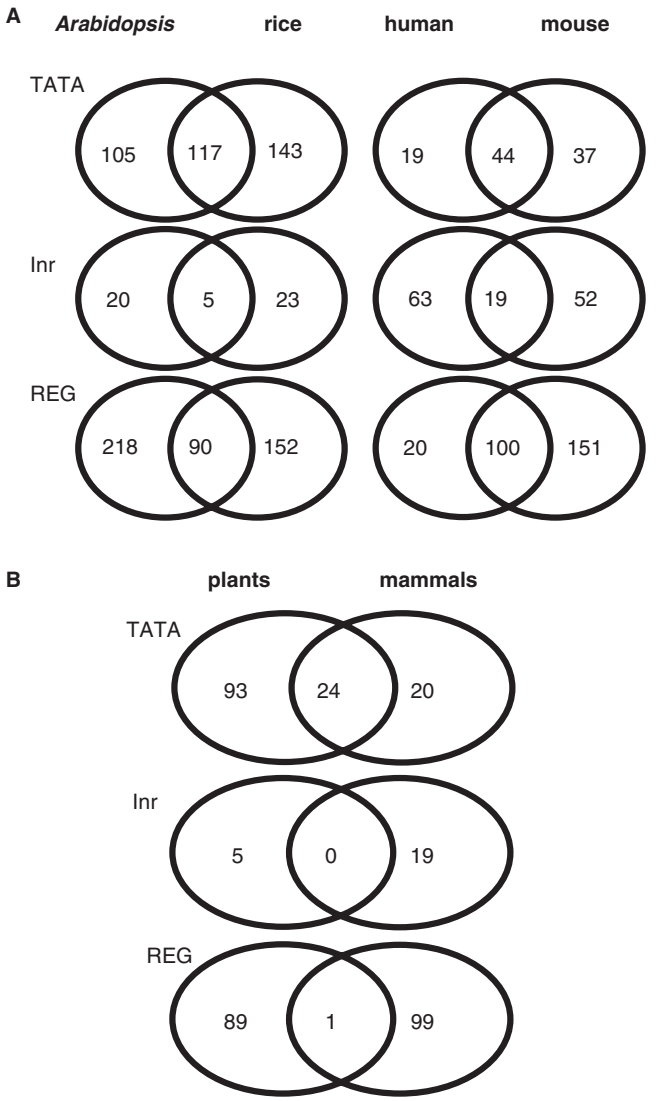

Figure 2. Conservation of octamer sequences. Numbers indicate conserved and differentiated octamer sequences. Sequences for plant TATA and REG were obtained from a previous report (26). Directional insensitivity was considered for the REG group. (**A**) Sequence conservation between *Arabidopsis* and rice, and between human and mouse. (**B**) Sequence conservation between plants and mammals. Sequence numbers of plants and mammals are of consensus sequences between *Arabidopsis*, and rice, and human and mouse, respectively, as shown in Panel A.

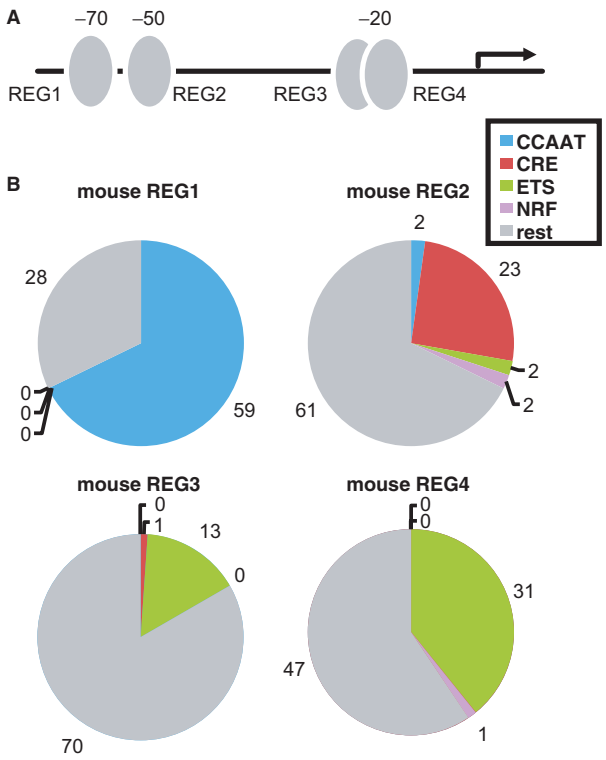

Figure 3. Mouse REG subgroups. Composition of motifs among mouse REG subgroups as shown in Figure 2. Numbering of the subgroups is the same as Figure 2. (**A**) Illustration of rough positioning of each REG subgroup. (**B**) Motif composition of each subgroup. See Table S6 for the sequence list and motif definition.

Recently, Blanchette *et al.* reported that distinct sets of regulatory elements are found according to where they appear: long range (>100 kb upstream), distal (1–10 kb upstream), proximal (0–1 kb upstream), downstream (1 kb from 1st intron), and intron type (intron) (7). Our results indicate that the proximal group can be further classified into subgroups according to their local distribution.

Absence of Sp1 in plant promoters

The Sp1 element of mammalian promoters is recognized by a transcription factor, Sp1 (35). Groups of mammalian octamers are associated with the element, but no plant group could be assigned to it (Figure 1 and Table 1). Further investigation of distribution profiles for the Sp1 core sequence revealed that the Sp1 element is not associated with plant promoters at all (Figure 4A), strongly suggesting that it is not used in plants. Supporting this idea, genes for the Sp1 factor are not found in the *Arabidopsis* or rice genomes (data not shown). Therefore, the absence of the Sp1 element in plant promoters is reasonable.

Absence of CpG-islands in plant promoters

Mammalian promoters are known to be associated with CpG-islands (1,36). It has been estimated that half of the mammalian promoters for protein-coding genes contain CpG islands (1). They are thought to have a role in core promoter function, especially in TATA-less promoters (1,4), but their exact function in transcription remains

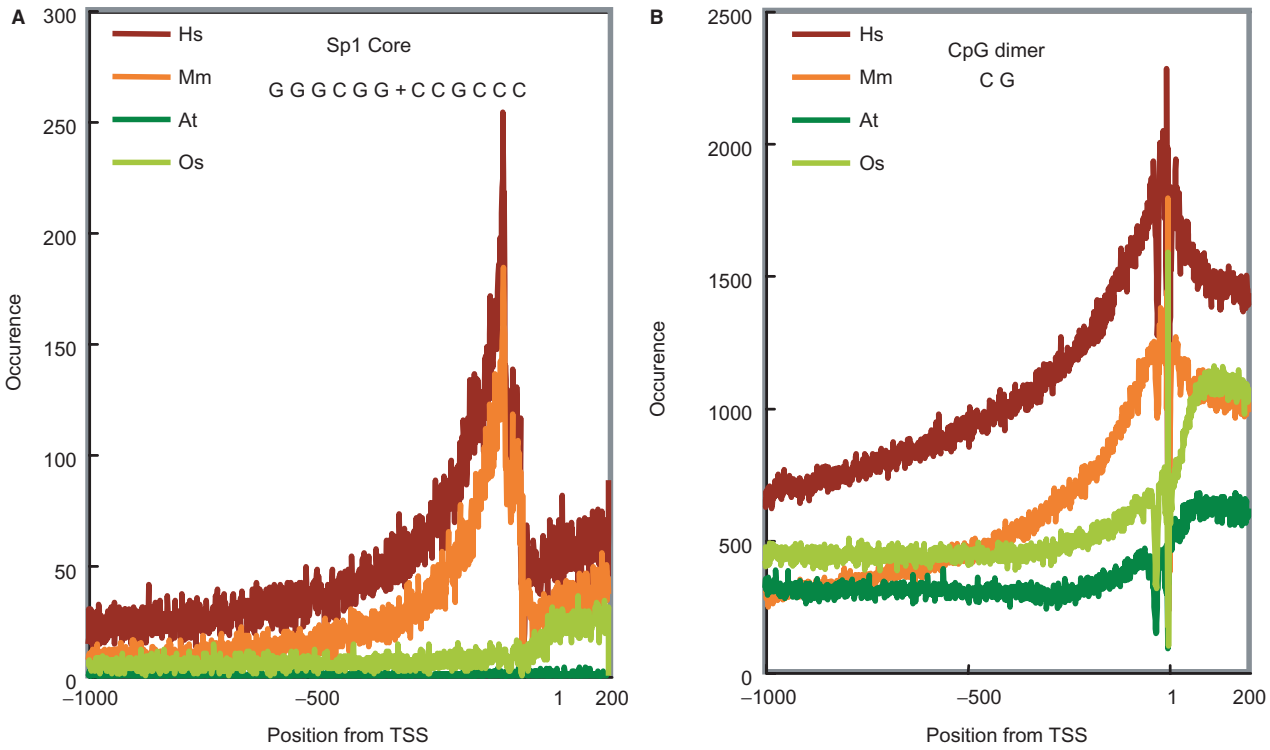

Figure 4. Absence of Sp1 and CpG-islands in plant promoters. (**A**) Distribution profiles of the core sequence of the Sp1 element [GGGCGG and its complementary sequence, CCGCCC (30)] are indicated in the promoter sets of human (Hs), mouse (Mm), *Arabidopsis* (At) and rice (Os). (**B**). Distribution profiles of the CpG dimer are shown.

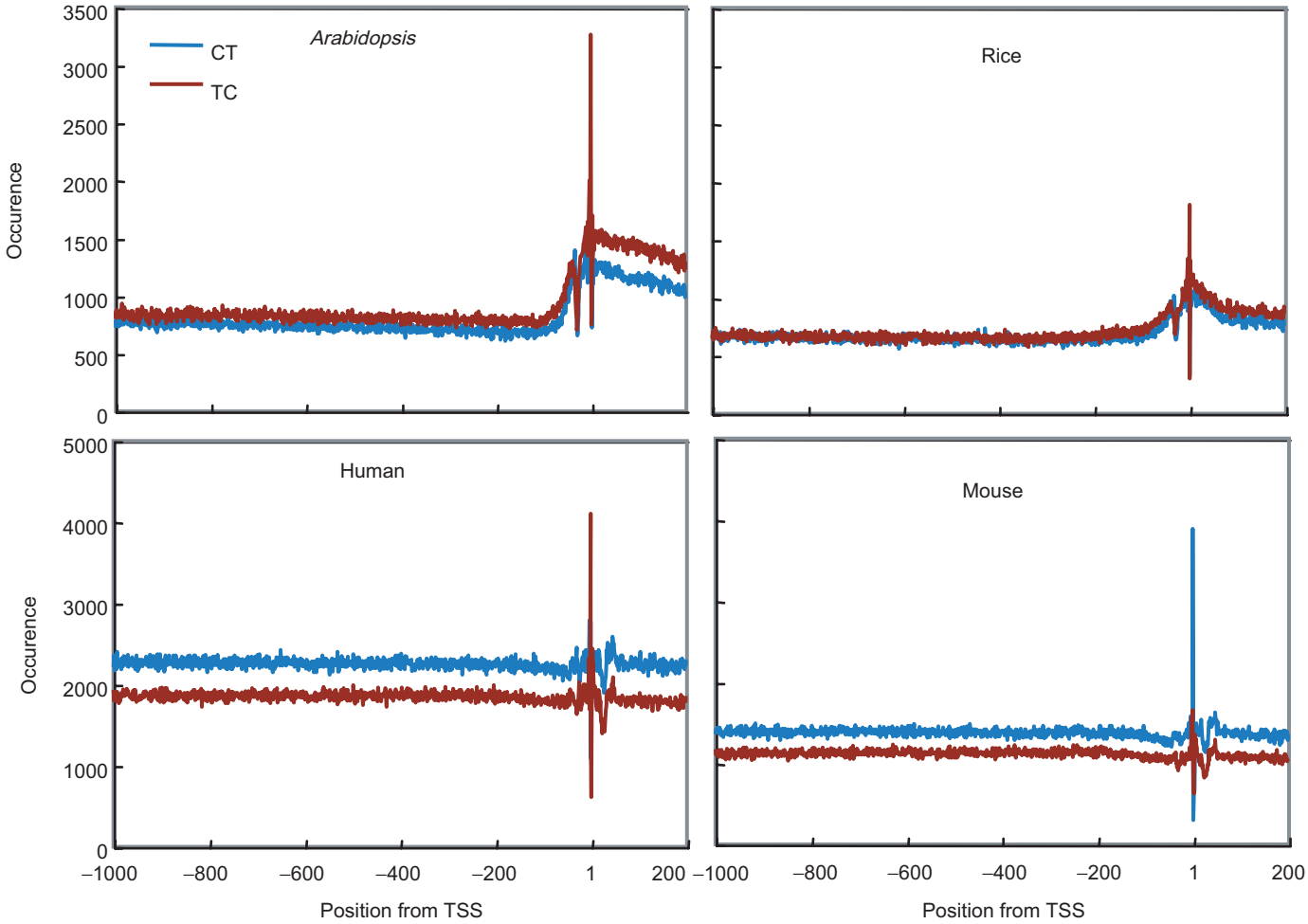

Figure 5. Absence of Y Patch in mammalian promoters. Distribution profiles of CT and TC dimers are shown. Sharp peaks correspond to Inr or Inr-like sequences.

poorly defined (1). It is also known that CpG islands are a target for epigenetic regulation and are supposed to achieve transcriptional control (37,38).

The CpG-related octamer group was detected in mammalian promoters by LDSS analysis (Figure 1 and Table S3 for sequences). However, it was not found in plant octamer groups (Figure 1 and summarized in Table 1). This observation raised the idea that plant promoters are not associated with CpG islands. In order to confirm this possibility, we subsequently examined distribution profiles of the CpG dimmer sequence along the promoter region (Figure 4B). As expected, human and mouse showed a distributional peak at the TSS, consistent with a previous study of human promoters (21). Our preliminary analysis suggested that the CpG dimmer is also associated with zebrafish TSS, though less drastic than mammals (Yamamoto and Obokata, unpublished data). In clear contrast, neither *Arabidopsis* nor rice promoters have a concentration of CpG dimmers at the TSS region, indicating that plant promoters are not associated with CpG islands. There are more CpG dimmers downstream of the TSS in *Arabidopsis* and rice, possibly due to the GC-rich protein-coding regions. These analyses have revealed that core promoter architecture being differentiated between mammals and higher plants. This is the first report to show such differentiation.

Mammalian 'broad type' promoters, where the TSSs are scattered evenly within a TSS cluster, are known to be associated with CpG islands (3,4). *Arabidopsis* also has 'broad type' promoters (Yamamoto and Obokata, unpublished data), but they are not associated with CpG islands.

Y Patch as a core promoter element specific to higher plants

Although there is a lack of CpG-related groups in plant octamers, plants have another group with a similar distribution profile, Y Patch (Figure 1). A typical Y Patch is composed of C and T (pyrimidine) (26) (Table S2). Our octamer LDSS analyses detected Y Patch only from *Arabidopsis* and rice and not from human and mouse (Table 1). These results indicate that Y Patch is a higher plant-specific element. In order to confirm this, we analyzed distribution profiles of pyrimidine dimers, CT and TC, for more sensitive detection of pyrimidine stretch (Figure 5). The results clearly shows

that mammalian promoters are not associated with pyrimidine stretch at all, in contrast to the results of *Arabidopsis* and rice promoters. Similar analyses for zebrafish promoters suggested absence of Y Patch in zebrafish promoters as well (Yamamoto and Obokata, unpublished data).

Both the mammalian CpG-related group and the plant Y Patch group have distribution peaks around the TSS and show a gradual decrease with increased distance from the TSS (Figure 1). However, we found no other shared characteristics between CpG islands and Y Patch. Different from the case of CpG islands, no methylation target is known for Y Patch. In addition, Y Patch has strict direction sensitivity in contrast to CpG-related groups (Table 1). Therefore, even though both groups have similar distribution profiles, Y Patch would not be a functional equivalent of a CpG island. Methylated cytosines have the potential to be mutated to thymines, thus m^5CpG has an evolutional tendency to be changed to TpG (39). Taking this into account, there is still no relationship between CpG and Y Patch.

Y Patch is detected not only by our LDSS analysis. A search for consensus sequence from *Arabidopsis* core promoters by MEME and AlignACE also discovered a Y Patch-related motif [Motif 1: TTTCTTCTTC, (40). *Arabidopsis* TSS regions are reported to show CG-compositional strand bias, or CG skew, where C is more frequently observed in the (+) strand than G (41,42). We suggest that this CG skew is a reflection of the presence of Y Patch around TSSs. According to the distribution profile and direction sensitivity of Y Patch, it has the potential to determine the direction of transcription, but its function is yet to be elucidated.

CONCLUSIONS

In this study, we have identified conservation and differentiation of core promoter architectures between higher plants and mammals by LDSS analysis. Comparison of *Arabidopsis*, rice, human and mouse promoters has revealed three categories: shared groups found in both higher plant and mammals (TATA box), common groups found in both kingdoms but the utilized sequence is differentiated (REG, Inr and Kozak), and specific groups found in either higher plants or mammals (CpG and Y Patch). The finding of the third group indicates differentiated architectures of higher plant and mammalian core promoters.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported in part by KAKENHI (Grant-in-Aid for Scientific Research) on Priority Areas 'Comparative Genomics' (to Y.Y.Y. and J.O.) and by Grant-in-Aid for Scientific Research (C) (to Y.Y.Y.) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. Funding to pay the Open Access publication charges for this article was provided by Grant-in-Aid for Scientific Research (C).

Conflict of interest statement. None declared.

REFERENCES

1. Smale,S.T. and Kadonaga,J.T. (2003) The RNA polymerase II core promoter. *Annu. Rev. Biochem.*, **72**, 449–479.
2. Butler,J.E. and Kadonaga,J.T. (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.*, **16**, 2583–2592.
3. Suzuki,Y., Tsunoda,T., Sese,J., Taira,H., Mizushima-Sugano,J., Hata,H., Ota,T., Isogai,T., Tanaka,T. *et al.* (2001) Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res.*, **11**, 677–684.
4. Carninci,P., Sandelin,A., Lenhard,B., Katayama,S., Shimokawa,K., Ponjavic,J., Semple,C.A., Taylor,M.S., Engstrom,P.G. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**, 626–635.
5. Taylor,M.S., Kai,C., Kawai,J., Carninci,P., Hayashizaki,Y. and Semple,C.A. (2006) Heterotachy in mammalian promoter evolution. *PLoS Genet.*, **2**, e30.
6. Cooper,S.J., Trinklein,N.D., Anton,E.D., Nguyen,L. and Myers,R.M. (2006) Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.*, **16**, 1–10.
7. Blanchette,M., Bataille,A.R., Chen,X., Poitras,C., Laganiere,J., Lefebvre,C., Deblois,G., Giguere,V., Ferretti,V. *et al.* (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.*, **16**, 656–668.
8. Lettice,L.A., Horikoshi,T., Heaney,S.J., van Baren,M.J., van der Linde,H.C., Breedveld,G.J., Joosse,M., Akarsu,N., Oostra,B.A. *et al.* (2002) Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proc. Natl Acad. Sci. USA*, **99**, 7548–7553.
9. Carter,D., Chakalova,L., Osborne,C.S., Dai,Y.F. and Fraser,P. (2002) Long-range chromatin regulatory interactions in vivo. *Nat. Genet.*, **32**, 623–626.
10. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
11. Roth,F.P., Hughes,J.D., Estep,P.W. and Church,G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.
12. Bailey,T.L. and Elkan,C. (1995) The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 21–29.
13. van Helden,J., Andre,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827–842.
14. Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.*, **296**, 1205–1214.
15. Marino-Ramirez,L., Spouge,J.L., Kanga,G.C. and Landsman,D. (2004) Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Res.*, **32**, 949–958.
16. Manson McGuire,A. and Church,G.M. (2000) Predicting regulons and their cis-regulatory motifs by comparative genomics. *Nucleic Acids Res.*, **28**, 4523–4530.
17. Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
18. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.B., Reynolds,D.B. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

19. Prakash,A. and Tompa,M. (2005) Discovery of regulatory elements in vertebrates through comparative genomics. *Nat. Biotechnol.*, **23**, 1249–1256.
20. Elkon,R., Linhart,C., Sharan,R., Shamir,R. and Shiloh,Y. (2003) Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res.*, **13**, 773–780.
21. FitzGerald,P.C., Shlyakhtenko,A., Mir,A.A. and Vinson,C. (2004) Clustering of DNA sequences in human promoters. *Genome Res.*, **14**, 1562–1574.
22. Shi,W. and Zhou,W. (2006) Frequency distribution of TATA Box and extension sequences on human promoters. *BMC Bioinformatics*, **7(Suppl. 4),** S2.
23. Tharakaraman,K., Marino-Ramirez,L., Sheetlin,S., Landsman,D. and Spouge,J.L. (2005) Alignments anchored on genomic landmarks can aid in the identification of regulatory elements. *Bioinformatics*, **21(Suppl. 1),** i440–i448.
24. Bajic,V.B., Tan,S.L., Suzuki,Y. and Sugano,S. (2004) Promoter prediction analysis on the whole human genome. *Nat. Biotechnol.*, **22**, 1467–1473.
25. Gershenzon,N.I., Trifonov,E.N. and Ioshikhes,I.P. (2006) The features of Drosophila core promoters revealed by statistical analysis. *BMC Genomics*, **7**, 161.
26. Yamamoto,Y.Y., Ichida,H., Matsui,M., Obokata,J., Sakurai,T., Satou,M., Seki,M., Shinozaki,K. and Abe,T. (2007) Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics*, **8**, 67.
27. Seki,M., Narusaka,M., Kamiya,A., Ishida,J., Satou,M., Sakurai,T., Nakajima,M., Enju,A., Akiyama,K. *et al.* (2002) Functional annotation of a full-length Arabidopsis cDNA collection. *Science*, **296**, 141–145.
28. Yamashita,R., Suzuki,Y., Wakaguri,H., Tsuritani,K., Nakai,K. and Sugano,S. (2006) DBTSS: Database of human transcription start sites, progress report 2006. *Nucleic Acids Res.*, **34**, D86–D89.
29. Mantovani,R. (1999) The molecular biology of the CCAAT-binding factor NF-Y. *Gene*, **239**, 15–27.
30. Sharrocks,A.D. (2001) The ETS-domain transcription factor family. *Nat. Rev. Mol. Cell. Biol.*, **2**, 827–837.
31. Scarpulla,R.C. (2002) Transcriptional activators and coactivators in the nuclear control of mitochondrial function in mammalian cells. *Gene*, **286**, 81–89.
32. Kiran,K., Ansari,S.A., Srivastava,R., Lodhi,N., Chaturvedi,C.P., Sawant,S.V. and Tuli,R. (2006) The TATA-box sequence in the basal promoter contributes to determining light-dependent gene expression in plants. *Plant Physiol.*, **142**, 364–376.
33. Juo,Z.S., Chiu,T.K., Leiberman,P.M., Baikalov,I., Berk,A.J. and Dickerson,R.E. (1996) How proteins recognize the TATA box. *J. Mol. Biol.*, **261**, 239–254.
34. Riechmann,J.L., Heard,J., Martin,G., Reuber,L., Jiang,C., Keddie,J., Adam,L., Pineda,O., Ratcliffe,O.J. *et al.* (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.
35. Kriwacki,R.W., Schultz,S.C., Steitz,T.A. and Caradonna,J.P. (1992) Sequence-specific recognition of DNA by zinc-finger peptides derived from the transcription factor Sp1. *Proc. Natl Acad. Sci. USA*, **89**, 9759–9763.
36. Ioshikhes,I.P. and Zhang,M.Q. (2000) Large-scale human promoter mapping using CpG islands. *Nat. Genet.*, **26**, 61–63.
37. Bird,A.P. and Wolffe,A.P. (1999) Methylation-induced repression – belts, braces, and chromatin. *Cell*, **99**, 451–454.
38. Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.
39. Millar,C.B., Guy,J., Sansom,O.J., Selfridge,J., MacDougall,E., Hendrich,B., Keightley,P.D., Bishop,S.M., Clarke,A.R. *et al.* (2002) Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice. *Science*, **297**, 403–405.
40. Molina,C. and Grotewold,E. (2005) Genome wide analysis of Arabidopsis core promoters. *BMC Genomics*, **6**, 25.
41. Tatarinova,T., Brover,V., Troukhan,M. and Alexandrov,N. (2003) Skew in CG content near the transcription start site in Arabidopsis thaliana. *Bioinformatics*, **19(Suppl. 1),** i313–i314.
42. Fujimori,S., Washio,T. and Tomita,M. (2005) GC-compositional strand bias around transcription start sites in plants and fungi. *BMC Genomics*, **6**, 26.

Table S1. Tobacco TATA activity and *Arabidopsis* TATA octamers

	GUS activity[1]	Peak position	RPH	*p* value
Prototype	5.00 ± 1.10			
TCACTATATATAG				
!TCACTATA		**-39**	**18.2**	**6.8×10^{-10}**
! CACTATAT		**-38**	**25.3**	**1.0×10^{-13}**
! ACTATATA		**-36**	**16.8**	**$<10^{-17}$**
! CTATATAT		**-35**	**32.2**	**$<10^{-17}$**
! TATATATA		**-34**	**13.3**	**$<10^{-17}$**
! ATATATAG		**-33**	**14.0**	**1.0×10^{-14}**
T_7>A				
******A******	1.91 ± 0.40			
TCACTAAATATAG				
TCACTAAA		-426	4.5	0.60
CACTAAAT		-309	5.7	0.80
ACTAAATA		-388	4.0	0.11
CTAAATAT		-315	4.2	0.58
! TAAATATA		**-32**	**7.2**	**2.1×10^{-8}**
AAATATAG		-111	3.5	0.21
T_7>C				
******C******	0			
TCACTACATATAG				
TCACTACA		-76	5.4	0.38
CACTACAT		-164	5.3	0.93
ACTACATA		-449	7.2	0.24
! CTACATAT		-37	7.8	9.8×10^{-3}
! TACATATA		-36	4.3	3.0×10^{-3}
ACATATAG		-989	5.8	0.37
T_7>G				
******G******	0			
TCACTAGATATAG				
TCACTAGA		-289	5.6	0.43
CACTAGAT		-295	7.0	0.72
ACTAGATA		-561	6.8	8.9×10^{-2}
CTAGATAT		-453	6.7	0.17
TAGATATA		-519	4.6	0.17
AGATATAG		-90	5.3	0.78
A_8>C				
*******C*****	0			
TCACTATCTATAG				

TCACTATC		-703	7.7	0.57
CACTATCT		-62	9.9	4.0×10^{-3}
ACTATCTA		-571	6.5	0.62
! CTATCTAT		-38	7.9	2.8×10^{-2}
! TATCTATA		-37	5.4	1.1×10^{-3}
ATCTATAG		-310	5.5	0.15
A_8>G				
*******G*****	0			
TCACTATGTATAG				
TCACTATG		-182	5.3	0.94
CACTATGT		-696	7.8	0.53
ACTATGTA		-309	4.2	0.34
CTATGTAT		-399	6.5	2.1×10^{-3}
! TATGTATA		-38	4.5	2.1×10^{-3}
ATGTATAG		-252	5.6	0.75
A_8>T				
*******T*****	1.85 ± 0.39			
TCACTATTTATAG				
TCACTATT		-647	5.3	0.55
CACTATTT		-352	4.0	0.90
! ACTATTTA		-34	4.8	2.7×10^{-2}
! **CTATTTAT**		**-35**	**13.4**	**1.4×10^{-10}**
! **TATTTATA**		**-33**	**8.9**	**4.9×10^{-11}**
! ATTTATAG		-32	5.2	5.9×10^{-2}

[1]GUS activity in the light (x 10^3 pmole MU min^{-1} mg $protein^{-1}$), data from (26). An exclamation mark means that the peak position is within the range of a TATA box. *p* value less than 10^{-4} (25) is emphasized in bold.

Functionality of *Arabidopsis* TATA octamers identified by LDSS analysis

Kiran *et al* reported the effects of extensive point mutations of a typical TATA element on transcriptional activity measured by tobacco transient assays (26). With the data presented in the report, we examined the relationship between distribution profiles of octamers and functional activity. According to the systematic mutation analysis, the 7th and 8th positions (TCACTATATAG) of the utilized 13-bp element have been found to be critical for TATA-dependent expression.

When distribution in the *Arabidopsis* promoter set of octamer components of the 13-bp TATA element was examined, all the 6 octamer components were revealed to be

TATA-related sequences, according to the peak positions (exclamation marks). In addition, two parameters for localized distribution, the RPH (Relative Peak Height, peak height over baseline) and the *p* value (a probability of the distribution under the assumption of random populations (25)) showed strong values of localization, indicating that these sequences are utilized as TATA elements with high specificity in the *Arabidopsis* genome.

As shown in the table, mutation of T_7>C, T_7>G, A_8>C, and A_8>G abolished the reporter expression, and these mutations sweep peak positions of the constituting octamer out of the range of the TATA region (T_7>G), or significantly increase *p* values (T_7>C, A_8>C, and A_8>G), an indication of a reduction of localized distribution. On the other hand, mutations with detectable reporter activity (T_7>A and A_8>T) still keep precise peak position and low *p* values (bold elements). These results clearly demonstrate functional implication of locally distributed elements.