Published online 18 October 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D977–D981***
doi:10.1093/nar/gkm785

ppdb: a plant promoter database

Yoshiharu Y. Yamamoto and Junichi Obokata*

Center for Gene Research, Nagoya University, Nagoya 464-8602, Japan

Received August 14, 2007; Revised September 13, 2007; Accepted September 17, 2007

ABSTRACT

ppdb (http://www.ppdb.gene.nagoya-u.ac.jp) is a plant promoter database that provides promoter annotation of *Arabidopsis* and rice. The database contains information on promoter structures, transcription start sites (TSSs) that have been identified from full-length cDNA clones and also a vast amount of TSS tag data. In ppdb, the promoter structures are determined by sets of promoter elements identified by a position-sensitive extraction method called local distribution of short sequences (LDSS). By using this database, the core promoter structure, the presence of regulatory elements and the distribution of TSS clusters can be identified. Although no differentiation of promoter architecture among plant species has been reported, there is some divergence of utilized sequences for promoter elements. Therefore, ppdb is based on species-specific sets of promoter elements, rather than on general motifs for multiple species. Each regulatory sequence is hyperlinked to literary information, a PLACE entry served by a plant *cis*-element database, and a list of promoters containing the regulatory sequence.

BACKGROUND

A promoter database can be generated from a combination of genome sequence, information of promoter positions and a list of *cis*-regulatory elements. Currently a major restriction on the quality of a promoter database is our limited knowledge of *cis*-elements. There are several established genome-wide plant promoter databases available today (RARGE: (1), http://www.rarge.gsc.riken.jp/; AGRIS: (2), http://www.arabidopsis.med.ohio-state.edu/; AthaMap: (3), http://www.athamap.de/), and which are based on *cis*-regulatory sequences from PlantCARE [(4), http://www.bioinformatics.psb.ugent.be/webtools/plantcare/html/], PLACE [(5), http://www.dna.affrc.go.jp/PLACE/] or TRANSFAC [(6), http://www.gene-regulation.com/pub/databases.html]. These three promoter databases focus on *cis*-regulatory elements rather than core promoter structure, aiming to reveal the regulatory machinery that give the expression profiles. Unfortunately, these databases provide information only for *Arabidopsis*, and there are no genome-wide plant promoter databases available for other plant species.

Local distribution of short sequences (LDSS) analysis is a method to extract promoter constituents by genome-wide statistical analysis (7,8). We have applied this method to the *Arabidopsis* and rice genomes, and identified 1000 octamer sequences per genome as LDSS-positive promoter elements (8). According to their distribution profiles, the identified octamers have been classified into regulatory element group (REG), TATA box and Y Patch as three major promoter element groups. REG is a direction-insensitive element that is preferentially found around −100 bp relative to the major transcription start site (TSS), and contains many established *cis*-regulatory sequences. Y Patch is a direction-sensitive plant core promoter element that appears around TSS. We found that utilized sequences of all three groups, including TATA element, are moderately differentiated between *Arabidopsis* and rice, demonstrating the importance of individual preparation of promoter elements for each genome.

The large collection of extracted promoter elements can be utilized as a tool to reveal precise promoter architecture. Therefore, here we present a novel searchable ppdb database, based on the LDSS-positive elements. Utilization of a genome-specific set of promoter elements and the detection of the core promoter structure are the two unique features of this database. Currently, ppdb is the only one plant promoter database with information about core promoter types on a genomic scale, and the first genome-wide database for rice promoters.

PROMOTER SELECTION AND OUTPUT WINDOWS

Major function of ppdb is to detect promoter elements in the genome sequence and to summarize promoter structures. Data source of ppdb is shown in Table 1. The database detects REG, TATA box and Y Patch.

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

Table 1. Source of ppdb

	Specification	Source	Size
Arabidopsis			
Genome sequence and gene annotation	TAIR 6	http://www.arabidopsis.org/	
TSS information	Cap signatured CT-MSS tags	Yamamoto, Y. Y. *et al.*, unpublished data	158 237
	Selected RAFL cDNA	http://rarge.gsc.riken.jp/	62 108
Promoter elements	LDSS-positive octamers	(8)	659
	PLACE entries corresponding to LDSS elements	http://www.dna.affrc.go.jp/PLACE/	21 (only matched motifs)
Rice			
Genome sequence and gene annotation	RGSP build 4.0	http://rapdb.lab.nig.ac.jp/	
TSS information	Selected fl cDNA	http://cdna01.dna.affrc.go.jp/cDNA/	17 286
Promoter elements	LDSS-positive octamers	(8)	600
	PLACE entries corresponding to LDSS elements	http://www.dna.affrc.go.jp/PLACE/	4 (only matched motifs)

Promoters of interest can be identified by a word search (e.g. 'photosystem') or a gene number (e.g. At5g38410, Os01g0100700 or AK121523) on the front page (http://www.ppdb.gene.nagoya-u.ac.jp). Selection of a specific gene model gives the following information: (i) sequence data, (ii) TSS data, (iii) a summary of the core promoter structure and (iv) REG data (Figure 1).

At the sequence window, octamer elements identified by the LDSS analysis (8) are highlighted. There are two modes for detection, 'Reliable' and 'ALL. Reliable' is a default setting where only elements at appropriate positions relative to the peak TSS are detected. Promoters without any TSS information do not show any elements. In this case, selecting *ALL* allows global detection without any positional restriction. The sensitive area in the *Reliable* mode for each element group is described on the front page.

The 'TSS information' table provides the expressional strength of each TSS. Tag per million (TPM) in the window shows the relative counts of TSS tags in a tag library, and this information comes from CT-MPSS analysis (Yamamoto,Y.Y. *et al.*, unpublished data). The methods and quality assessment of the data will be described elsewhere.

'The table of Core promoter information' shows the presence and absence of TATA box and Y Patch. Currently, a search for Inr (Initiator for the consensus around TSS) is not executed, thus all promoters will show 'Not Available' for it. We have a plan to add Inr information in a near future as a minor update.

The 'REG information' table shows a REG list of a promoter and its corresponding PPDB and PLACE motifs. For example, the table in Figure 1 shows that AtREG379 belongs to the ACGT group of PPDB and corresponds to ACGT, ACGTG, GCCAC and ACGTGKC motifs of PLACE (5). PPDB motifs have been extracted from REG sequences with the aid of a two-dimensional (2D) REG-promoter clustering (8). REG sequences, as well as PPDB and PLACE motifs, are linked to other pages for biological information.

REG information is shown under the category of 'Promoter Summary'. Selection of *ALL* adds another category, 'Not Reliable Promoter Summary'. This category can be used when searching for regulatory elements (REG) from wider regions or when there is no TSS information on the promoter of interest.

ADDITIONAL PAGES

Some biological information of REGs is also provided by ppdb. As shown in Figure 2, there is a page with a whole list of REGs, the 'ALL REG List'. This list presents the relationship between REG sequences, PPDB motifs and PLACE motifs.

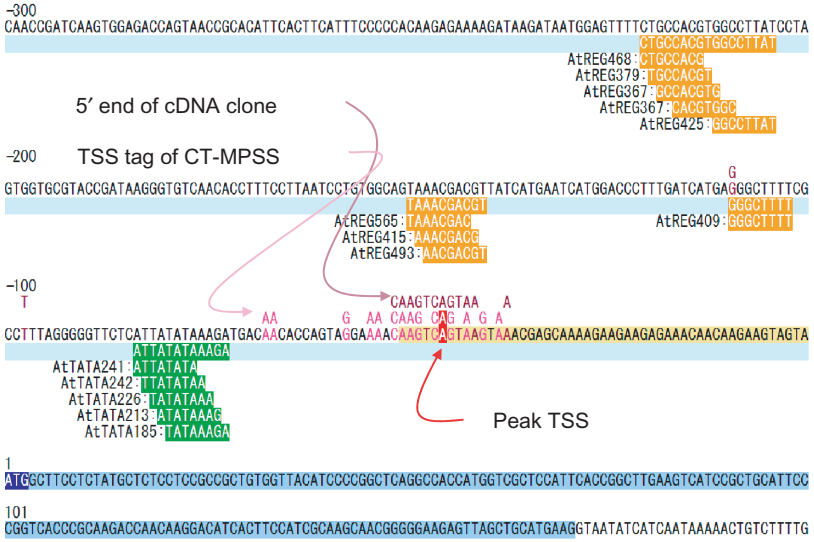
Selection of a specific REG sequence leads to 'Summary of the REG', followed by 'Hit Gene List of the REG'. The 'Summary of the REG' section shows corresponding PPDB and PLACE motifs and a brief description of the motifs. Selection of each motif leads to PLACE ID and also to the original article(s) for the motif. The 'Hit Gene List' section gives information about promoters sharing the REG.

ACKNOWLEDGEMENTS

This work was supported by Grant-in-Aid for Scientific Research on Priority Areas 'Comparative Genomics' (Y.Y.Y. and J.O.), Scientific Research (B) (J.O.) and Scientific Research (C) (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 Ministry of Education, Culture, Sports, Science and Technology, Japan.

Conflict of interest statement. None declared.

Promoter Sequence



- **Promoter Summary of AT5G38410.1**

TSS information:

TSS	Sequence	TPM score	Genome position	Position from initiation codon
TSS	A	2	Chromosome 5(−) 15395604	-68
TSS	A	5	Chromosome 5(−) 15395603	-67
TSS	G	5	Chromosome 5(−) 15395594	-58
TSS	A	2	Chromosome 5(−) 15395591	-55
TSS	A	10	Chromosome 5(−) 15395590	-54
TSS	C	5	Chromosome 5(−) 15395588	-52
TSS	A	1041	Chromosome 5(−) 15395587	-51
TSS	A	92	Chromosome 5(−) 15395586	-50
TSS	G	213	Chromosome 5(−) 15395585	-49
TSS	C	15	Chromosome 5(−) 15395583	-47
TSS peak	A	2982	Chromosome 5(−) 15395582	-46
TSS	G	51	Chromosome 5(−) 15395581	-45
TSS	A	844	Chromosome 5(−) 15395579	-43
TSS	G	8	Chromosome 5(−) 15395577	-41
TSS	A	209	Chromosome 5(−) 15395575	-39

Core promoter information:

Type	Sequence	Genome position	Position from initiation codon
initiator	Not Available	Not Available	Not Available
TATA Box	ATTATATAAAGA	Chromosome 5(−) 15395609–15395620	-84 – -73
Y Patch	None	None	None

REG information:

REG	Sequence	Genome position		Position from initiation codon	
REG	GGGCTTTT	Chromosome 5(−) 15395639–15395646		-110 – -103	
AtREG409	GGGCTTTT	PPDB Motif	GCCCA	PLACE Motif	
REG	TAAACGACGT	Chromosome 5(−) 15395677–15395686		-150 – -141	
AtREG565	TAAACGAC	PPDB Motif		PLACE Motif	
AtREG415	AAACGACG	PPDB Motif		PLACE Motif	
AtREG493	AACGACGT	PPDB Motif	ACGT	PLACE Motif	ACGT
REG	CTGCCACGTGGCCTTAT	Chromosome 5(−) 15395741–15395757		-221 – -205	
AtREG468	CTGCCACG	PPDB Motif		PLACE Motif	GCCAC
AtREG379	TGCCACGT	PPDB Motif	ACGT	PLACE Motif	ACGT, ACGTG, GCCAC, ACGTGKC
AtREG367	CACGTGGC	PPDB Motif	ACGT	PLACE Motif	ACGT, ACGTG, GCCAC, CACGTG, ACGTGKC, CACGTGGC
AtREG425	GGCCTTAT	PPDB Motif	AACCG(G/A)	PLACE Motif	

Figure 1. Selection of a specific gene model gives the following information: (i) sequence data, (ii) TSs data, (iii) a summary of the core promoter structure and (iv) REG data. Peak TSS is highlighted. TPM means tag per million and this is an indication of expression level at each TSS.

- **All REG List (Total: 308)**

ID	Sequence	PPDB Motif	PLACE Motif
AtREG351	GGGCCTTA	GCCCA	GGGCC
AtREG352	ATTGGGCC	GCCCA	GGGCC, TGGGCY
AtREG353	AAGGCCCA	GCCCA	GGGCC, TGGGCY
AtREG354	AGGCCCAT	GCCCA	GGGCC, TGGGCY
AtREG355	GCCCAATA	GCCCA	
AtREG356	AGGCCCAA	GCCCA	GGGCC, TGGGCY
AtREG357	GCCCATTA	GCCCA	
AtREG358	TAGGCCCA	GCCCA	GGGCC, TGGGCY
AtREG359	AAAGGCCC	GCCCA	GGGCC
AtREG360	GGGCCTAA	GCCCA	GGGCC
AtREG361	AATGGGCC	GCCCA	GGGCC, TGGGCY
AtREG362	ATAGGCCC	GCCCA	GGGCC
AtREG363	CTGGGCCC	GCCCA	GGGCC, TGGGCY
AtREG364	GGCCCATA	GCCCA	GGGCC, TGGGCY
AtREG365	GGGCTTTA	GCCCA	
AtREG366	CACGTGTC	ACGT	ACGT, ACGTG, CACGTG, ACGTGKC, ACACNNG, ACGTGTC

- **Summary of AtREG378 (All REG)**

ID	AtREG378	
Sequence	AAGCCCAT	
PPDB Motif	GCCCA	Element II of Arabidopsis PCNA-2, cell cycle/meristematic expression
PLACE Motif	TGGGCY	"Site II element" found in the promoter regions of cytochrome genes (Cytc-1, Cytc-2) in Arabidopsis; Located between -147 and -156 from the translational starts sites (Welchen et al., 2005); Y=C/T; See also S000308; Overrepresented in the promoters of nuclear genes encoding components of the oxidative phosphorylation (OxPhos) machinery from both Arabidopsis and rice (Welchen and Gonzalez, 2006);)

- **Hit Gene list of AtREG378 (Total: 1143)**

Locus	Gene model	Sequence	Description
AT1G75560	AT1G755601	TAAAAGCCCAT	zinc knuckle (CCHC-type) family protein, contains Pfam domain, PF00098: Zinc knuckle
AT1G17210	AT1G172101	TAAAGCCCATTTA	expressed protein, distantly related to dentin phosphoryn (Homo sapiens) (GI:4322670)
AT1G77770	AT1G777702	TAAATGGGCTTACGCCCCATTAT	expressed protein
AT1G29965	AT1G299651	AAAAGGCCCAATGGGCTTTA	60S ribosomal protein L18A (RPL18aA), JRW
AT1G04010	AT1G040101	CAAAGGCCCAATTAAAAGCCCATTT	lecithin:cholesterol acyltransferase family protein / LACT family protein, weak similarity to SP|P40345 Phospholipid:diacylglycerol acyltransferase (EC 2.3.1.158) (PDAT) (Saccharomyces cerevisiae); contains Pfam profile PF02450: Lecithin:cholesterol acyltransferase (phosphatidylcholine-sterol acyltransferase)
AT1G63855	AT1G638553	AAATGGGCTTAGGCC	expressed protein
AT1G68590	AT1G685901	CAAAGCCCATAA	plastid-specific 30S ribosomal protein 3, putative / PSRP-3, putative, similar to SP|P82412 Plastid-specific 30S ribosomal protein 3, chloroplast precursor (PSRP-3) (Spinacia oleracea); contains Pfam profile PF04839: Plastid and cyanobacterial ribosomal protein (PSRP-3 / Ycf65)
AT1G68590	AT1G685901	ATGGGCTTAT	plastid-specific 30S ribosomal protein 3, putative / PSRP-3, putative, similar to SP|P82412 Plastid-specific 30S ribosomal protein 3, chloroplast precursor (PSRP-3) (Spinacia oleracea); contains Pfam profile PF04839: Plastid and cyanobacterial ribosomal protein (PSRP-3 / Ycf65)

Figure 2. Page showing the whole list of REGs, the 'All REG List', 'Summary of the REG' followed by 'Hit Gene List of the REG'. This also shows the relationship between REG Sequences, PPDB motifs and PLACE motifs.

REFERENCES

1. 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.
2. Davuluri,R.V., Sun,H., Palaniswamy,S.K., Matthews,N., Molina,C., Kurtz,M. and Grotewold,E. (2003) AGRIS: *Arabidopsis* gene regulatory information server, an information resource of *Arabidopsis cis*-regulatory elements and transcription factors. *BMC Bioinformatics*, **4**, 25.
3. Bülow,L., Steffens,N.O., Galuschka,C., Shindler,M. and Hehl,R. (2006) AthaMap: from *in silico* data to real transcription factor binding sites. *In Silico Biol.*, **6**, 0023.
4. Lescot,M., Déhais,P., Thijs,G., Marchal,K., Moreau,Y., Van de Peer,Y., Rouzé,P. and Rombauts,S. (2002) PlantCARE, a database

of plant *cis*-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. *Nucleic Acids Res.*, **30**, 325–327.
5. Higo,K., Ugawa,Y., Iwamoto,M. and Korenaga,T. (1999) Plant *cis*-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.*, **27**, 297–300.
6. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M. *et al*. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
7. FitzGerald,P.C., Shlyakhtenko,A., Mir,A.A. and Vinson,C. (2004) Clustering of DNA sequences in human promoters. *Genome Res.*, **14**, 1562–1574.
8. 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.