Proceedings of ANAS (Biological and Medical Sciences), vol. 73, No 2, p. 33-41 (2018)

The PlantProm: A Database of Plant Promoter Sequences (Release 2016)

I.A. Shahmuradov[1,2*], A.U. Abdulazimova[2], M. Genaev[3] and V.V. Solovyev[3]

[1] Bioinformatics Laboratory, Institute of Molecular Biology & Biotechnologies, Azerbaijan National Academy of Sciences, 2A Matbuat Ave., Baku AZ1073, Azerbaijan;
[2] Institute of Biophysics, Azerbaijan National Academy of Sciences;
[3] Softberry Inc. (USA)
**E-mail: ilhambaku@gmail.com*

Knowledge on promoter sequences and their characteristics is crucial for improving our basic understanding of gene regulation. In 2003, we launched the PlantProm database of 305 plant proximal promoter sequences for RNA polymerase II with experimentally determined transcription start site (TSS). Here, we present a new release of the PlantProm database that contains 576 entries including 150, 403 and 23 promoters of monocot, dicot and other plant genes, respectively, as well as high-throughput annotated and predicted promoters for five plant genomes. The database provides DNA sequences of promoters and their taxonomic/promoter type classification, occurrence of sequence motifs of known plant transcription factor binding sites in promoters, Nucleotide Frequency Matrices for two important promoter elements as TATA-box and Initiator element. In addition, the database includes computationally predicted TSS for 22,257 genes of *Oryza sativa*, 23,334 genes of *Zea mays*, 18,226 genes of *Medicago truncatula*, 38,702 genes of *Glycine max* and 11,037 genes of *Vitis vinifera*. The PlantProm DB is publicly available on http://www.softberry.com/plantprom2016/.

Keywords: *RNA polymerase II, plant promoter, transcription start site, database, promoter elements*

INTRODUCTION

Promoters occupy genomic regions upstream of and around transcription start site (TSS). Information on promoter sequences is fundamental for interpreting gene expression patterns, and constructing and understanding genetic regulatory networks. Transcription factor (TF) binding sites (TFBSs) that define specificity and rate of transcription are positioned in both proximal and distal promoter regions; TFBSs responsible for TSS selection are mostly localized in the proximal promoter, within a hundred nucleotides around the TSS (for review see: Solovyev et al., 2010; Hernandez-Garcia and Finer, 2014; Roy and Singer, 2015). To date, we are still far from complete understanding of genome architecture and functions. Experimental and computational approaches to this problem face significant challenges such as: (a) the mechanisms determining transcriptional status of gene(s) and choice of TSS are still mostly unclear and depend on cell/tissue type, developmental stage and environmental signals (Verona et al., 2008; Zou et al., 2008); (b) Experimental identification of TSSs is still quite expensive and time-consuming; (c) development of computational tools for predicting TSS(s) requires representative learning sets of experimentally validated promoters, but these data are still very limited (Hernandez-Garcia and Finer, 2014; Roy and Singer, 2015).

There are two types of available plant promoter collections: (1) Sets of promoters with TSS(s) determined by the genome-wide mapping of full-length cDNAs (FL-cDNA) and/or 5’-end tagging approaches, as CAGE, 5’-SAGE and TEC-RED (for review see Harbers and Carninci, 2005), presented in plant promoter databases (DB) such as RARGE DB (Sakurai et al., 2005; Akiyama et al., 2014) and ppdb (Yamamoto and Obokata, 2008; Hieno et al., 2014). In particular, the FL-cDNA technology provides valuable information on transcriptional units and facilitates identification of TSSs (Seki et al., 2002; Kikuchi et al., 2003; Ogihara et al., 2004; Sato et al., 2009; Soderlund et al., 2009; Matsumoto et al., 2011; Fukami-Kobayashi et al., 2014). (2) Sets of promoters with TSS(s) identified by direct experimental approaches, as the primer extension assay (Carey et al., 2013) and 5’-RACE (Rapid Amplification of cDNA Ends) assays (Scotto–Lavino et al., 2006), collected in Eukaryotic Promoter Database (EPD; Dreos et al., 2013, 2015) and PlantProm DB (Shahmuradov et al., 2003). EPD was the first representative collection of eukaryotic RNA polymerase II (Pol II) promoters with TSS(s) identified by direct experimental approaches (Praz et al., 2002). However, human and animal promoters prevail in this collection. Promoters of only two plant species, *Arabidopsis thaliana* and *Zea mays,* are currently represented in EPD (Dreos et al., 2013, 2015).

The latest release (version 3.0) of the ppdb

(Hieno et al., 2014; http://ppdb.agr.gifu-u.ac.jp/ppdb/cgi-bin/index.cgi) is the biggest source on TSS positions for plant species, providing information on experimentally mapped TSSs of four plant species, as Arabidopsis, rice, poplar and moss (*Physcomitrella patens*). In particular, the ppdb contains TSS information for all Arabidopsis (27,206) and 12,535 (out of 32,325) rice protein-coding genes annotated in these genomes. However, our analysis of these TSS positions relative to the start points of the annotated coding DNA sequences (CDS) indicates that in some cases the distance between TSSs and CDS start positions is less than 10 base pairs (bp). In particular, we revealed 7,878 (~29%) and 1,554 (~13%) such "TSS-CDS" pairs in Arabidopsis and rice, respectively. Although the minimum length of 5'-untranslated region (UTR) for mRNAs remains unknown, many studies conclude that 5'-UTR should be longer than 20 bp for the efficient binding of ribosomes and initiation of translation (Li and Wan, 2004; Chen et al., 2011; Kim et al., 2014; Hinnebusch et al., 2016). So, our findings indicate that some subset of TSSs collected in the ppdb remains to be verified in future studies.

With the development of advanced experimental techniques, significant progress has been made in the genome-wide identification of promoters/TSSs and analysis of gene regulatory sequences (for review see Mundade et al., 2014; Suryamohan and Halfon, 2015; Levati et al., 2016). Recently, Geng et al. (2014) developed a high-yield screening system in peanut by establishing a simple digital expression profile based on Illumina sequencing that allows, in particular, tissue-specific promoter cloning. However, TSSs identified by these techniques lie only approximately around the real start points of transcription and, therefore, remain to be verified by the other more precise methods such as 5'-RACE (Shiraki et al., 2003; Hashimoto et al., 2004). Therefore, such promoter collections are not suitable for retrieving position-specific promoter features adjacent to the TSS, which is often exploited in computational tools for TSS prediction. To date, the most accurate promoter prediction programs (e.g. see: Shahmuradov et al., 2005; Anwar et al., 2008) have been developed by using promoter sets from PlantProm DB and/or EPD databases that include experimentally verified exact TSS positions.

Previously, we developed PlantProm DB collecting 305 experimentally verified plant Pol II promoters from many published sources (Shahmuradov et al., 2003). It has been used to study a variety of plant biology problems, which include investigating differential expression of soluble pyrophosphatase isoforms in Arabidopsis (Oeztuerk et al., 2015), cis-regulatory elements in plant cell signaling (Priest et al., 2009), a functional role for DNA methylation in transcription (Aceituno et al., 2008) and transcription of nuclear organellar DNA in plants (Wang et al., 2014), as well as many studies of computational promoter identification (Shahmuradov et al., 2005; Pandey and Krishnamachari, 2006; Gan et al., 2009; Tatarinova et al., 2013). All these results demonstrate the importance of our promoter collection.

Here we present a new release of PlantProm DB with 576 experimentally verified promoter sequences, enlarging our collection of 305 promoters from the first release. We provide a structural classification of these promoters and Nucleotide Frequency Matrices (NFM) for their important functional elements, such as TATA box and Initiator element (INR). Applying TSSPlant promoter prediction program (see its description below), we performed the genome-wide search of putative TSSs for protein-coding genes from 5 plant species (*Oryza sativa*, *Z. mays*, *Medicago truncatula*, *Glycine max* and *Vitis vinifera*). Results of these studies are included in this release of the PlantProm DB. Moreover, the new release contains information on statistically significant motifs of 3,032 known plant TFBSs found in 576 experimentally verified promoter sequences and in [-1000:+101] promoter regions of 113,556 genes of 5 plant genomes. At last, we significantly improved the DB interface and its search capabilities.

METHODS

To collect plant promoters with TSS position validated by direct experiments, such as primer extension and 5'-RACE assays, we applied essentially the same rules as described previously (Shahmuradov et al., 2003). To select non-redundant promoter sequences, we used BLAST program (Altschul et al., 1997) for pairwise comparisons of [-50:+1] promoter regions and kept only promoters showing less than 90% sequence homology in these regions.

To classify promoter sequences into the TATA and TATA-less promoters, as well as to compute NFMs for TATA and INR elements, we applied the Expectation Maximization (EM) algorithm (Cardon and Stormo, 1992). Details of EM algorithm for this task were described previously (Shahmuradov et al., 2003).

To predict putative TSSs in genomic sequences we applied novel promoter prediction tool, TSSPlant (Shahmuradov et al., 2017). TSSPlant predicts both TATA and TATA-less promoters in sequences of wide spectrum of plant genomes. It demonstrated significantly higher accuracy compared to other known and available promoter prediction programs, including TSSP program, trained on previous version of PlantProm DB (http://www.softberry.com/berry.phtml). TSSPplant tool is now available for online running (http://www.softberry.com/berry.phtml?topic=tssplant&group=programs&subgroup=promoter).

For genome-wide search of putative promoters (TSSs) in higher plants we selected protein-coding genes of 5 species: monocots *O. sativa, japonica* (35,655 genes; genome assembly IRGSP-1.0) and *Z. mays* (36,988 genes; genome assembly AGPv3), dicots *M. truncatula* (47,202 genes; genome assembly MedtrA17_4.0), *G. max* (53,151 genes; genome assembly v1.0) and *V. vinifera* (26,118 genes; genome assembly IGGP_12x) from Ensembl genome browser annotation system (http://plants.ensembl.org/info/website/ftp/index.html). For promoter analysis only genes with annotated 5'-UTR length of 20 bp or more were selected. If the selected gene had several gene (mRNA) start points, we consider further only a variant with the longest 5'-UTR. For promoter search we extracted [-1000:+101] regions from the above selected genes, where +1 corresponds to the gene annotated start position. In total, we obtained [-1000:+101] regions for 22,332, 23,467, 18,227, 38,718 and 11,079 genes from *O. sativa, Z. mays*, *M. truncatula*, *G. max* and *V. vinifera*, respectively.

Search for statistically significant motifs of 3,032 known plant TFBSs from the Regsite database (http://www.softberry.com/berry.phtml?topic=regsite) was performed by Nsite program (Shahmuradov and Solovyev, 2015; see also: http://www.softberry.com/plantprom2016/). Nsite executes searches for statistically non-random motifs of known TFBSs in a single DNA sequence. A predicted motif is considered as statistically significant if (i) the expected (by chance) number of such motifs in a given nucleotide sequence is less than an assigned threshold and (ii) the total number of identified motifs is equal to or greater than the upper limit of 95% confidence interval. The search and statistical estimations are performed separately on both strands of a query sequence.

PlantProm database was implemented using Apache WEB Server running on CentOS Linux. MySQL was used as a server database. The server part of Web interface was written in PHP. Modules for downloading gff3 (general feature format 3) annotations and sequence files for individual promoters were written in Perl. The "Search services" used to retrieve information from data tables were implemented using JavaScript library.

RESULTS

General Structure and Content of the PlantProm DB style

Fig. 1 shows the structure and content of PlantProm DB. It consists of seven main modules:

(1) Promoters from direct experiments;
(2) Putative TSS map for protein-coding genes;
(3) Classification of promoters;
(4) Canonical NFMs;
(5) Nucleotide composition;
(6) Regulatory motifs;
(7) Search services.

PlantProm DB release 2016.03 is available at http://www.softberry.com/plantprom2016/. It provides user-friendly interface: all data can be retrieved and downloaded.

Promoters from direct experiments

The module "Promoters from direct experiments" allows a user to retrieve and download 576 promoter sequences of 251 bp length from 87 plant species with TSS identified by primer extension assay and/or 5'-RACE assays, where position 201 corresponds to the experimentally validated TSS (+1). The set includes 305 promoters from the first release and 271 newly added promoters. If this module is chosen in the Main Menu, the sub-menu displayed in Fig. 2 appears. Here, depending on chosen option ("view" or "download") for the selected set of promoters, a user can view or download promoter sequences in FASTA format; with the "view" option, TATA-boxes and transcribed regions are displayed in upper case.

The module "Classification of promoters" is composed of functions to retrieve and download various taxonomic and promoter type (TATA or TATA-less) classes of 576 promoters. It consists of two sections: "Summary" and "Individual Characteristics". In the first section, a list of all species represented in the experimentally verified promoter collection and data on the total number and the number of promoters' of each class are given for each species. If the user visits the "Individual Characteristics" section that is organized as a table, many individual characteristics of genes/promoters and original data sources, including GenBank and PubMed links for every annotated promoter, will be displayed (http://www.softberry.com/data/plantprom/Links/Taxon_Table_2.htm).

The module "Canonical NFMs" allows database users to retrieve and download TATA-box and INR NFMs for various classes of promoters.

The module "Nucleotide composition" contains data on nucleotide composition of promoter regions of various classes, including sequences before the TSS, [-200:-1], and after the TSS, [+1:+51]; the user can view and download this information.

TSSs in five model plant genomes

The module "Putative TSS map for protein-coding genes" allows the user to retrieve and download locations of putative TSSs predicted by TSSPlant program in [-1000:+101] regions of 113,556 protein-coding genes of five plant species (*O. sativa*, *Z. mays*, *M. truncatula*, *G. max* and *V. vinifera)*. In this module, for every genome, 4 options are given (Fig. 3). The user can view and download data on predicted TSSs for every gene in gff or text formats, get information on every gene (gene name and product, genomic positions of a gene and mRNA and CDS starts, number of alternative mRNAs, length of longest 5'-UTR, etc.) and view/download [-1000:+101] region in FASTA format.

Regulatory motifs

The module "Regulatory motifs" contains data on statistically significant (E-value < 0.01; for details of the statistical estimations see Shahmuradov and Solovyev, 2015) motifs of 3,032 known TFBSs and their consensuses in both experimentally verified promoters and [-1000:+101] regions of protein-coding genes from five plant species (Fig. 4). For experimentally verified promoters, the user can view these data for every promoter (out of 576). For 113,556 genes from five species, *O. sativa*, *Z. mays*, *M. truncatula*, *G. max* and *V. vinifera*, a single Nsite output file for every genome is supplied.

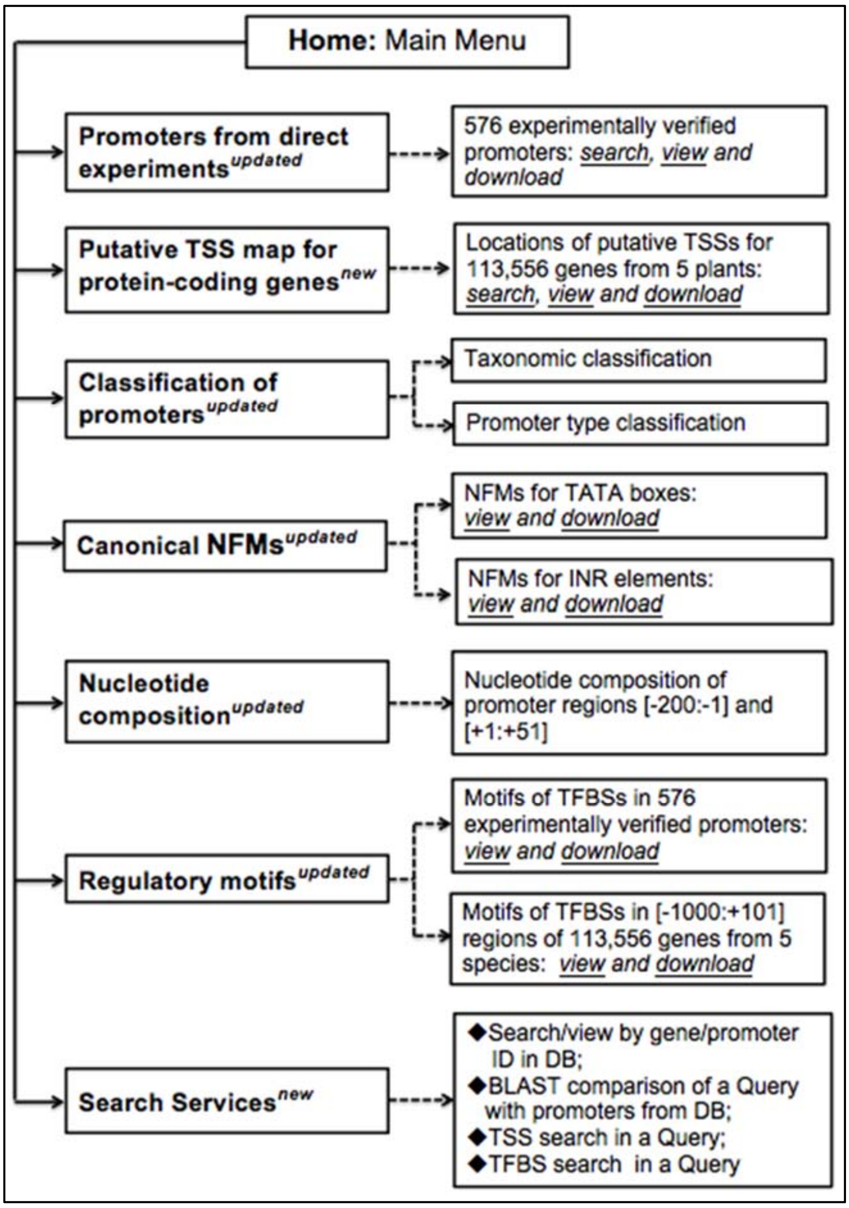
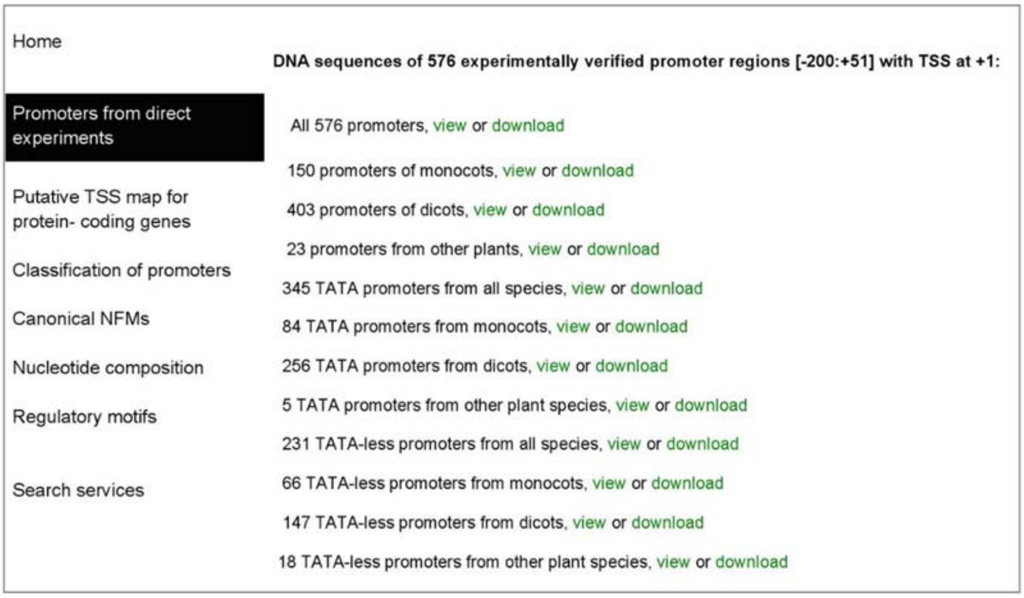

Home: Main Menu
Promoters from direct experiments^updated
576 experimentally verified promoters: search, view and download
Putative TSS map for protein-coding genes^new
Locations of putative TSSs for 113,556 genes from 5 plants: search, view and download
Classification of promoters^updated
Taxonomic classification
Promoter type classification
Canonical NFMs^updated
NFMs for TATA boxes: view and download
NFMs for INR elements: view and download
Nucleotide composition^updated
Nucleotide composition of promoter regions [-200:-1] and [+1:+51]
Regulatory motifs^updated
Motifs of TFBSs in 576 experimentally verified promoters: view and download
Motifs of TFBSs in [-1000:+101] regions of 113,556 genes from 5 species: view and download
Search Services^new
◆Search/view by gene/promoter ID in DB;
◆BLAST comparison of a Query with promoters from DB;
◆TSS search in a Query;
◆TFBS search in a Query

Fig. 1. The structure and content of PlantProm DB. New and significantly updated modules are marked ("*new*" or "*updated*").


Home
Promoters from direct experiments
Putative TSS map for protein- coding genes
Classification of promoters
Canonical NFMs
Nucleotide composition
Regulatory motifs
Search services

DNA sequences of 576 experimentally verified promoter regions [-200:+51] with TSS at +1:

All 576 promoters, view or download

150 promoters of monocots, view or download

403 promoters of dicots, view or download

23 promoters from other plants, view or download

345 TATA promoters from all species, view or download

84 TATA promoters from monocots, view or download

256 TATA promoters from dicots, view or download

5 TATA promoters from other plant species, view or download

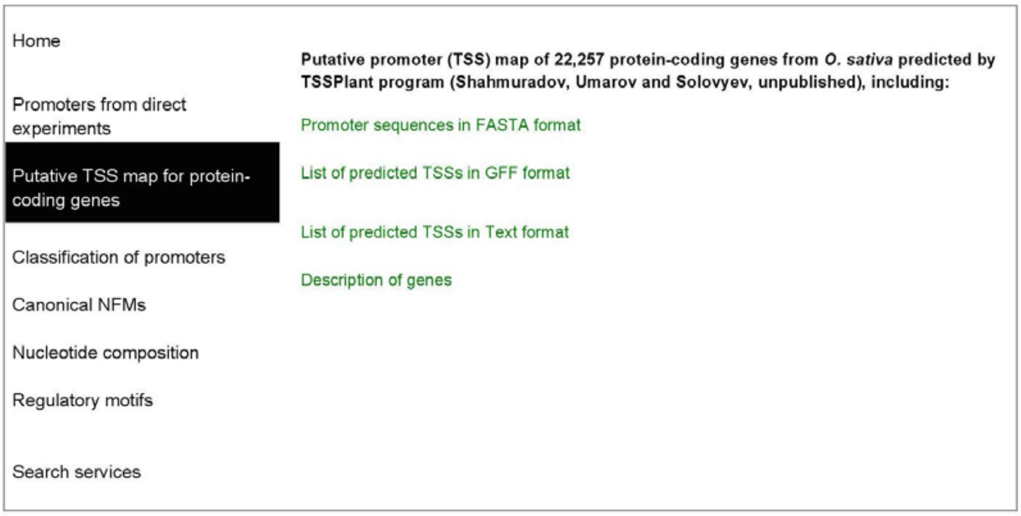
231 TATA-less promoters from all species, view or download

66 TATA-less promoters from monocots, view or download

147 TATA-less promoters from dicots, view or download

18 TATA-less promoters from other plant species, view or download

Fig. 2. The information content of the “Promoters from direct experiments” module.


Home
Promoters from direct experiments
Putative TSS map for protein-coding genes
Classification of promoters
Canonical NFMs
Nucleotide composition
Regulatory motifs
Search services

Putative promoter (TSS) map of 22,257 protein-coding genes from *O. sativa* predicted by TSSPlant program (Shahmuradov, Umarov and Solovyev, unpublished), including:

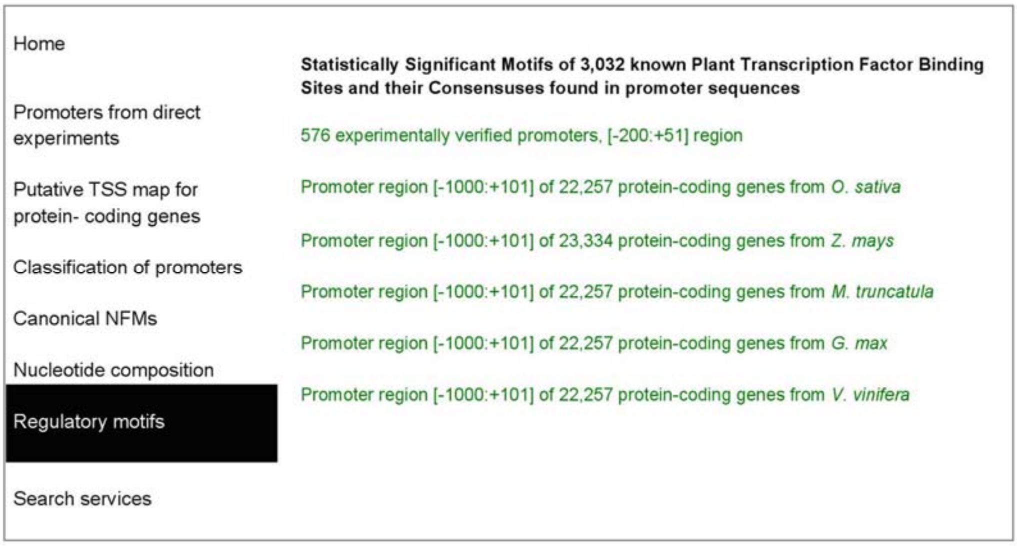
Promoter sequences in FASTA format

List of predicted TSSs in GFF format

List of predicted TSSs in Text format

Description of genes

Fig. 3. The informational content of the “Putative TSS map for protein-coding genes” module for *O. sativa* genome.


Home
Promoters from direct experiments
Putative TSS map for protein- coding genes
Classification of promoters
Canonical NFMs
Nucleotide composition
Regulatory motifs
Search services

Statistically Significant Motifs of 3,032 known Plant Transcription Factor Binding Sites and their Consensuses found in promoter sequences

576 experimentally verified promoters, [-200:+51] region

Promoter region [-1000:+101] of 22,257 protein-coding genes from *O. sativa*

Promoter region [-1000:+101] of 23,334 protein-coding genes from *Z. mays*

Promoter region [-1000:+101] of 22,257 protein-coding genes from *M. truncatula*

Promoter region [-1000:+101] of 22,257 protein-coding genes from *G. max*

Promoter region [-1000:+101] of 22,257 protein-coding genes from *V. vinifera*

Fig. 4. The informational content of “Regulatory motifs” module.

Search services

Utilizing five options of "Search services" module, the user can retrieve, view and download promoters by their promoter identifier (ID; in set of 576 promoters) or gene ID (in set of 113,556 genes from five species), as well as perform comparison of a query sequence with promoter sequences from PlantProm DB, search for TSS and motifs of 3032 known plant TFBSs.

Option "Search for promoters from direct experiments"

The promoters of interest can be selected (a) by checking their corresponding boxes on the left side of the WEB page or (b) by performing a search using a keyword. Afterwards, if "Get fasta" button is clicked, a page with sequences of selected promoters in FASTA format will appear for a view and downloading. Moreover, promoters can be sorted by the GenBank accession number, organism name, gene name and product.

Option "Search for putative TSS map for protein-coding genes"

For this option the same search and sorting rules are used, as in the case of "Search for promoters from direct experiments". However, here, the selected promoters can be viewed in two popular (FASTA and gff) formats.

Option "BLAST search"

If the user chooses this option, the BLAST program search window will appear. To perform the BLAST search, the following steps are required: (i) paste a query sequence in FASTA format or browse and select a file from your local folder; (ii) choose a promoter set from the given list; (iii) choose the alignment option (**Pairwise** or **Tabular**) and (iv) click **Process** button.

Option "Nsite tool"

When the user chooses this option, the window of search of TFBS motifs by Nsite program is displayed; here, a set of known plant transcription regulatory motifs can be searched in a query sequence.

3.5.5 Option "TSSPlant tool"

If users choose this option, the window of search of putative TSSs by TSSPlant program in a query sequence will appear.

DISCUSSION

The described new release of PlantProm DB contains enlarged collection of experimentally verified promoter sequences and includes several novel additions, such as descriptions of functional motifs in promoter sequences, the computational promoter annotations of five plant genomes, and improved retrieval and search possibilities for different promoter and genome characteristics. In particular, comparison of nucleotide composition of promoter sequences upstream and downstream of TSS in dicots and monocots revealed a significant difference between them in the promoter upstream regions: in dicots they are significantly more A/T-rich.

For 113,556 out of 113,823 genes (99.8%) from 5 genomes, at least one TSS was predictd by TSSPlant program. We computed a distribution of distances between a TSS described in the Ensembl genome annotation (TSSan) and the closest predicted TSS (TSSpr). Such distribution for *G. max* is shown in Fig. 2 (for other genomes see: Supplementary Fig. S7, S8, S9 and S10,). For 55,864 out of 108,938 genes (51.2%), one of the predicted TSSs is located relatively close (at a distance ≤50 bp) to the annotated start site of transcription. However, for ≃49% genes, the predicted TSSs are observed at larger distances from the annotated gene starts. Of course, some of such cases can be explained by a limited prediction capacity of TSSPlant, which is true for all promoter recognition tools published to date. Beyond this possibility, we can consider the followings. We analyzed protein-coding genes with annotated 5'-UTR longer than 20 bp. Among them, for 1,826, 1,218, 1,064, 1,178 and 1,897 genes from *O. sativa, Z. mays, G. max, M.truncatula* and *V. vinifera* genomes, respectively, the annotated length of the longest 5'-UTR was less than 40 bp. To date, the minimal length of 5'-UTR required for proper processing and translation of mRNA is unknown. However, in the same genomes, the longest mRNAs for 8,145, 11,333, 4,606, 17,149, 5,640, 5,238 and 2,828 genes have 5'-UTR lengths of 300 nucleotides or more. This observation can suggest that for significant portion of analyzed genes the annotated 5'-UTRs are truncated, and therefore the distance between the predicted TSS and actual gene start is shorter than we currently observe. Thus, if we take 100 bp (the approximate length of a typical core promoter; Roym and Singer, 2015) as acceptable maximum discrepancy between the predicted TSS and the annotated gene start, then TSSpr for 70,352 genes (~65%) is localized within that range. Another observation of our studies is that the total number of predicted TSSs per gene varies between 2 and 3. It partially agrees with ppdb data for rice: if we consider TSSs separated by 300 bp or more, two TSSs for 257 genes and three TSSs for 15 genes will be presented in the database. So, multiple TSSs seem to be a typical trait of the plant promoter architecture.

All high-throughput promoter identification approaches have their limitations in accuracy of promoter localization, so it is important to support a manually created database with high quality TSSs and promoter sequences derived from direct experimental studies of particular genes. At the same time, many genome annotation databases such as UCSC (Speir et al., 2016) and Ensembl (Yates et al., 2016) genome browsers contain experimentally discovered and predicted genes (from automatic annotations). It would be beneficial for various gene regulation studies to provide information on promoter location for each annotated gene, i.e. to add putative promoters derived by computational predictions to the current databases' content. We are currently preparing such information alongside with high-throughput promoter identification data for a set of sequenced plant genomes beyond the five already represented in this release.

PlantProm DB furnishes a representative learning set of promoter sequences that is essential for development of plant promoter prediction programs. Annotated regulatory motifs can be used for interpreting gene expression patterns and understanding genetic regulatory networks.

In animals (human, mice, Drosophila, etc.), many genes are regulated by multiple alternative promoters rather than a single promoter (Batut et al., 2013; Hernandez-Garcia and Finer, 2014). Study of alternative promoters has received little attention in plants, although recent advances in genomics and sequencing technologies would accelerate studies of alternate promoter usage in plants (Hernandez-Garcia and Finer, 2014). We are planning to update PlantProm DB regularly including available alternative promoter information.

FUNDING

This work has been supported by the grant from the Science Development Foundation under the President of the Republic of Azerbaijan [**EIF-2010-1(1)-40/27-3**].

REFERENCES

Aceituno F.F. et al. (2008) The rules of gene expression in plants: organ identity and gene body methylation are key factors for regulation of gene expression in *Arabidopsis thaliana*. *BMC Genomics*, **9:** 438.

Akiyama K. et al. (2014) RARGE II: an integrated phenotype database of Arabidopsis mutant traits using a controlled vocabulary. *Plant Cell Physiol.*, **55:** e4.

Altschul S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.*, **25:** 3389-3402.

Anwar F. et al. (2008) Pol II promoter prediction using characteristic 4-mer motifs: a machine learning approach. *BMC Bioinformatics*, **9:** 4.

Batut P. et al. (2013) High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res*., **23:** 169-180.

Cardon L., Stormo G. (1992) Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *J. Mol. Biol.*, **5:** 159-170.

Carey M.F. et al. (2013) The primer extension assay. *Cold Spring Harb. Protoc.*, **2013:** 164-173.

Chen C-H. et al. (2011) The genomic features that affect the lengths of 5' untranslated regions in multicellular eukaryotes. *BMC Bioinformatics*, **12(Suppl):** S3.

Dreos R. et al. (2013) EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucl. Acids Res.*, **41:** D157-D164.

Dreos R. et al. (2015) The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. *Nucl. Acids Res.*, **43:** D92-D96.

Fukami-Kobayashi K. et al. (2014) SABRE2: a database connecting plant EST/full-length cDNA clones with Arabidopsis information. *Plant Cell Physiol.*, **55:** e5.

GanY. et al. (2009)A pattern-based nearest neighbor search approach for promoter prediction using DNA structural profiles. *Bioinformatics*, **25:** 2006-2012.

Geng L. et al. (2014) Mining tissue-specific contigs from peanut (*Arachis hypogaea* L.) for promoter cloning by deep transcriptome sequencing. *Plant Cell Physiol.*, **55:** 1793-1801.

Harbers M., Carninci P. (2005) Tag-based approaches for transcriptome research and genome annotation. *Nature Methods*, **2:** 495-502.

HashimotoS. et al. (2004) 5'-end SAGE for the analysis of transcriptional start sites. *Nature Biotechnol.*, **22:** 1146-1149.

Hernandez-Garcia C.M., Finer J.J. (2014) Identification and validation of promoters and cis-acting regulatory elements. *Plant Sci.*, **217-218:** 109-119.

Hieno A. et al. (2014) ppdb: plant promoter database version 3.0. *Nucl. Acids Res.*, **42:** D1188–D1192.

Hinnebusch A.G. et al. (2016) Translational control by 5′-untranslated regions of eukaryotic mRNAs. *Science*, **352:** 1413-1416.

Kikuchi S. et al. (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science*, **301:** 376-379.

Kim Y. et al. (2014) The immediate upstream region of the 5′-UTR from the AUG start codon has a pronounced effect on the translational efficiency in

Arabidopsis thaliana. *Nucl .Acids Res.*, **42:** 485-498.

Li L., Wan C.C. (2004) Capped mRNA with a Single Nucleotide Leader Is Optimally Translated in a Primitive Eukaryote, *Giardia lamblia*. *J. Biol. Chem.*, **279:** 14656-14664.

Matsumoto T. et al. (2011) Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries. *Plant Physiol.*, **156:** 20–28.

Mundade R. et al. (2014) Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. *Cell Cycle*, **13:** 2847-2852.

Natsume S. et al. (2015) The Draft Genome of Hop (*Humulus lupulus*), an Essence for Brewing. *Plant Cell Physiol.*, **56:** 428-441.

Oeztuerk Z.N. et al. (2015) A Differential expression of soluble pyrophosphatase isoforms in Arabidopsis upon external stimuli. *Turk. J. Bot.*, **39:** 571–579.

Ogihara Y. et al. (2004) Construction of a full-length cDNA library from young spikelets of hexaploid wheat and its characterization by large-scale sequencing of expressed sequence tags. *Genes Genet. Syst.*, **79:** 227–232.

Pandey S.P., Krishnamachari A. (2006) Computational analysis of plant RNA Pol-II promoters. *Biosystems*, **83:** 38-50.

Priest H.D. et al. (2009) cis-regulatory elements in plant cell signaling. *Cur. Opin. Plant Biol.*, **12:** 643–649.

Roym A.L., Singer D.S. (2015) Core promoters in transcription: old problem, new insights. *Trends Biochem. Sci.*, **40:** 165–171.

Sakurai T. et al. (2005) RARGE: a large-scale database of RIKEN Arabidopsis resources ranging from transcriptome to phenome. *Nucl. Acids Res.*, **33:** D647–D650.

Sato K. et al. (2009) Development of 5006 full-length CDNAs in barley: a tool for accessing cereal genomics resources. *DNA Res.*, **16:** 81–89.

Scotto-Lavino E. et al. (2006) 5'end cDNA amplification using classic RACE. *Nature Protocols*, **1**, 2555–2562.

Seki M. et al. (2002) Functional annotation of a full-length Arabidopsis cDNA collection. *Science*, **296:** 141–145.

Shahmuradov I.A. et al. (2003) PlantProm: a database of plant promoter sequences. *Nucl. Acids. Res.*, **31:** 114–117.

Shahmuradov I.A., Solovyev V.V. (2015) Nsite, NsiteH and NsiteM computer tools for studying transcription regulatory elements. *Bioinformatics*, **31**, 3544–3545.

Shahmuradov I.A. et al. (2005) Plant promoter prediction with confidence estimation. *Nucl. Acids Res.*, **33:** 1069–1076.

Shahmuradov I.A., Umarov R.Kh., Solovyev V.V. (2017) TSSPlant: a new tool for prediction of plant Pol II promoters. Nucleic Acids Res**., 45(8):** e65.

Shiraki T. et al. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. USA*, **100:** 15776-15781.

Soderlund C. et al. (2009) Sequencing, mapping, and analysis of 27,455 maize full-length cDNAs. *PLoS Genet.*, **5:** e1000740.

Solovyev V.V. et al. (2010) Identification of promoter regions and regulatory sites. In: *Computational Biology of Transcription Factor Binding (Methods in Molecular Biology)*. Editor: Istvan Ladunga. Springer Science+Business Media, Humana Press, **674(Chapter 5):** 57-83; DOI 10.1007/978-1-60761-854-6_5.

Speir M.L. et al. (2016) The UCSC Genome Browser database: 2016 update. *Nucl. Acids Res.*, **44:** D717–D725.

Suryamohan K., Halfon M.S. (2015) Identifying transcriptional cis-regulatory modules in animal genomes. *Wiley Interdiscip. Rev. Dev. Biol.*, **4:** 59-84.

Tatarinova T. et al. (2013) NPEST: a nonparametric method and a database for transcription start site prediction. *Quant. Biol.*, **1:** 261-271.

Verona R.I. et al. (2008) The transcriptional status but not the imprinting control region determines allele-specific histone modifications at the imprinted H19 locus. *Mol. Cell Biol.*, **28:** 71-82.

Wang D. et al. (2014) Transcription of nuclear organellar DNA in a model plant system. *Genome Biol. Evol.*, **6:** 1327-1334.

Yamamoto Y.Y. et al. (2007) Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis. *Nucl. Acids Res.*, **35:** 6219–6226.

Yamamoto Y.Y., Obokata J. (2008) ppdb: a plant promoter database. *Nucl Acids Res.*, **33:** D977–D981.

Yates A. et al. (2016) Ensembl 2016. *Nucl. Acids Res.*, **44:** D710-D716.

Zou X. et al. (2008) Human glycolipid transfer protein (GLTP) genes: organization, transcriptional status and evolution. *BMC Genomics*, **9:** 72.

PlantProm: Bitki Promotor Ardıcıllıqları Üzrə Verilənlər Bazası (Buraxılış 2016)

İ. Ə. Şahmuradov[1,2*], Ə. Ü. Abduləzimova[2], M.Genayev[3] və V.V. Solovyev[3]

[1] AMEA Molekulyar Biologiya və Biotexnologiyalar İnstitutunun Bioinformatika laboratoriyası
[2]AMEA Biofizika İnstitutu
[3] Softberry Şirkəti (ABŞ)

Promotor ardıcıllıqları və onlar səciyyəvi xüsusiyyətləri haqqında biliklər gen tənzimlənməsinin əsaslarının başa düşülməsi üçün həlledici əhəmiyyət kəsb edir. 2003-cü ildə biz RNA polimeraza II üçün transkripsiya start saytı (TSS) təcrübi yolla müəyyənləşdirilmiş 305 bitki proksimal promotor ardıcıllığı üzrə PlantProm verilənlər bazasını təqdim etmişdik. Bu işdə biz PlantProm verilənlər bazasının yeni buraxılışını təqdim edirik. Həmin bazaya birləpəli, ikiləpəli və digər bitkilərdən müvafiq surətdə 150, 403 və 23 promotordan ibarət 576 nümunə, həmçinin 5 bitki genomunun annotasiya olunmuş və güman olunan promotorları üzrə məlumatlar daxildir. Verilənlər bazasında promotorların DNT ardıcıllıqları və onların taksonomik/promotor sinifləri üzrə təsnifatı, promotorlarda transkripsiya faktorlarının birləşmə saytları, TATA-boks və Initiator kimi 2 mühüm promotor elementi üzrə nukleotid tezlikləri matrisləri verilir. Bundan əlavə, verilənlər bazasına *Oryza sativa, Zea mays, Medicago truncatula, Glycine max* və *Vitis vinifera* bitkilərinin müvafiq surətdə 22257, 23334, 18226, 38702 11037 geni üçün potensial TSS-lər üzrəməlumatlar daxildir. PlantProm vürilənlər bazası http://www.softberry.com/plantprom2016/ səhifəsində mövcuddur.

Açar sözlər: *RNT polyimeraza II, bitki promotoru, transkripsiya start saytı, verilənlər bazası, promotor elementləri*

PlantProm: База Данных по Промоторным Последовательностям Растений (Выпуск 2016)

И. А. Шахмурадов[1,2], А. У. Абдулазимова[2], М.Генаев[3] и В. В. Соловьев[3]

[1] Лаборатория биоинформатики Института молекулярной биологии и биотехнологий НАН Азербайджана
[2] Институт биофизики НАН Азербайджана
[3] Softberry Inc. (США)

Знания о последовательностях промотора и их характеристиках имеет решающее значение для понимания основ регуляции генов. В 2003 году мы представили базу данных PlantProm по 305 проксимальным промоторным последовательностям растений для РНК-полимеразы II с экспериментально выявленным сайтом старта транскрипции (ССТ). Здесь мы представляем новую версию базы данных PlantProm, которая включает 576 записей, включая 150, 403 и 23 промотора генов однодольных, двудольных и других растений, соответственно, а также аннотированные и предсказанные промоторы для пяти геномов растений. В базе данных представлены последовательности ДНК промоторов и их классификация по таксономическим/промоторным классам, последовательности мотивов известных сайтов связывания факторов транскрипции растений в промоторах, матрицы нуклеотидных частот для элементов TATA-bокс и *Initiator*. Кроме того, база данных включает в себя предсказанные ССТ для 22257 генов *Oryza sativa*, 23334 гена *Zea mays*, 18226 генов *Medicago truncatula*, 38 702 гена *Glycine max* и 11 037 генов *Vitis vinifera*. База данных PlantProm доступна на http://www.softberry.com/plantprom2016/.

Ключевые слова: *РНК полимераза II, промоторы растений, сайт старта транскрипции, база данных, промоторные элементы*