Get Started Using PlantProm DB

The current release PlantProm DB allows users to:

(1) Retrieve and download 576 experimentally verified promoter sequences, classified by promoter class and taxonomy;

(2) Retrieve and download in FASTA format promoter sequences and a putative TSS map, in both text and GFF formats, for 113,556 protein-coding genes of *O. sativa*, *Z. mays*, *M. truncatula*, *G. max* and *V. vinifera*.

(3) Get a PubMed link for every entry for 576 experimentally verified promoters;

(4) Retrieve and download TATA-box and INR NFMs.

(5) Get information on nucleotide composition of promoter regions before [-200:-1] and after [+1:+51] TSS in various sets of 576 experimentally verified promoters;

(6) Retrieve and download putative TFBS contents of 576 experimentally verified promoter sequences.

(7) Retrieve and download putative TFBS contents of [-1000:-101] regions of 22,257, 23,334, 18,226, 38,702 and 11,037 protein-coding genes of *O. sativa*, *Z. mays*, *M. truncatula*, *G. max* and *V. vinifera*, respectively.

(8) Search for promoter sequences by promoter or gene ID.

(9) Perform BLAST comparison of a user-given query sequence with both experimentally verified promoters and [-1000:+1] regions of protein-coding genes of *O. sativa*, *Z. mays*, *M. truncatula*, *G. max* and *V. vinifera*.

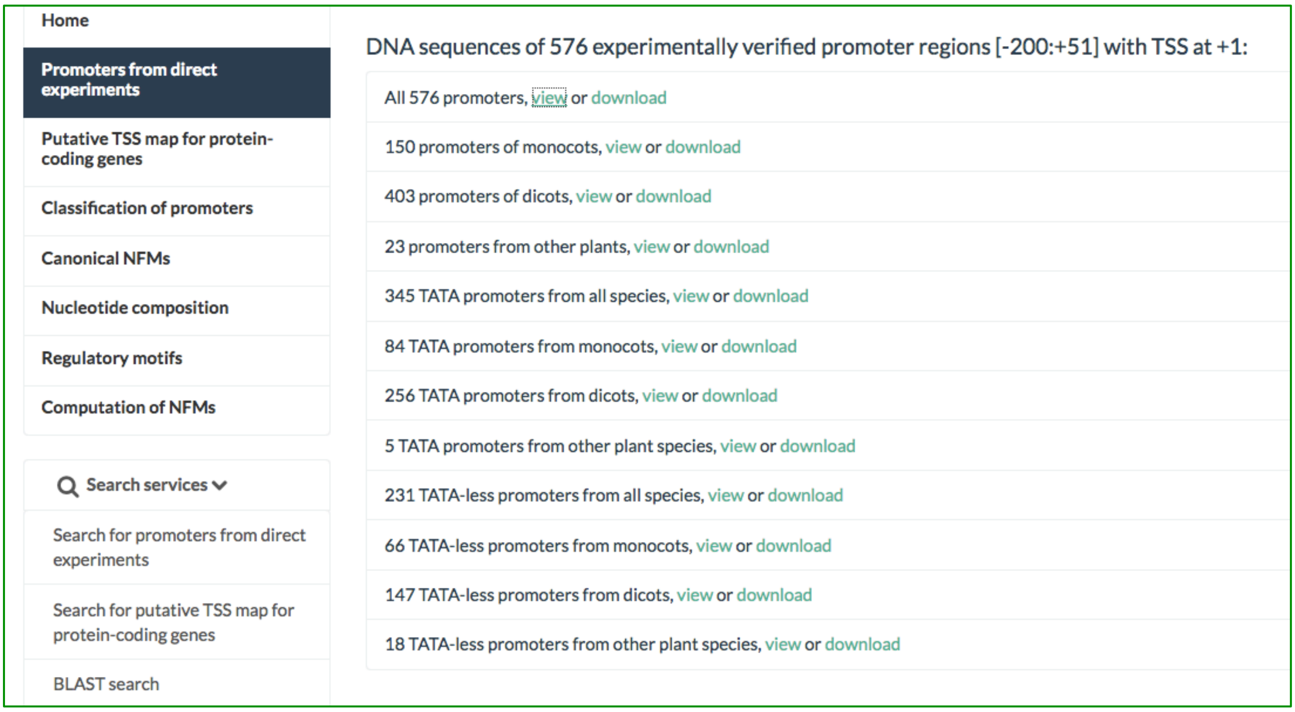
The menu and search service of PlantProm DB use JavaScript. Commonly used browsers - Internet Explorer, Firefox, Safari etc. – can be used with the database.

How to use PlantProm DB

Once you connect to PlantProm DB home page, you are ready to start working with DB.

View and download experimentally verified promoters

In Main Menu, if an option **“Promoters from direct experiments**” is chosen, the following sub-menu appears:


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

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

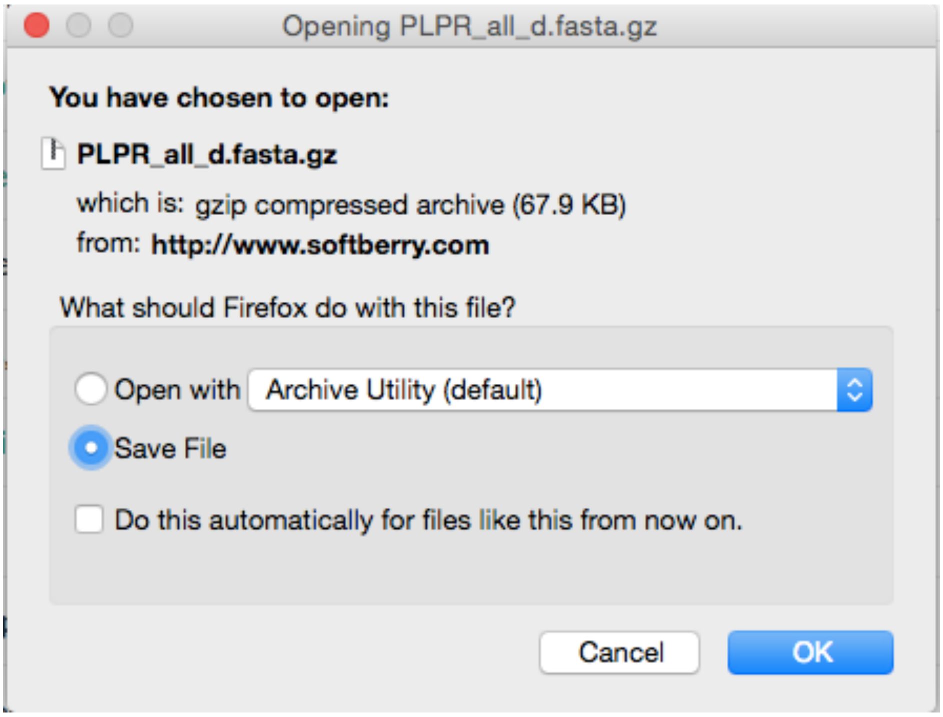
66 TATA-less promoters from monocots, view or download

147 TATA-less promoters from dicots, view or download

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

In this sub-menu, for a selected group of promoters,

- If an option **view** is clicked, the corresponding 251-bp promoter sequences in FASTA format are displayed;
- If an option **download** is clicked, the following prompt window is displayed:


Opening PLPR_all_d.fasta.gz

You have chosen to open:

PLPR_all_d.fasta.gz

which is: gzip compressed archive (67.9 KB)
from: http://www.softberry.com

What should Firefox do with this file?

○ Open with Archive Utility (default)

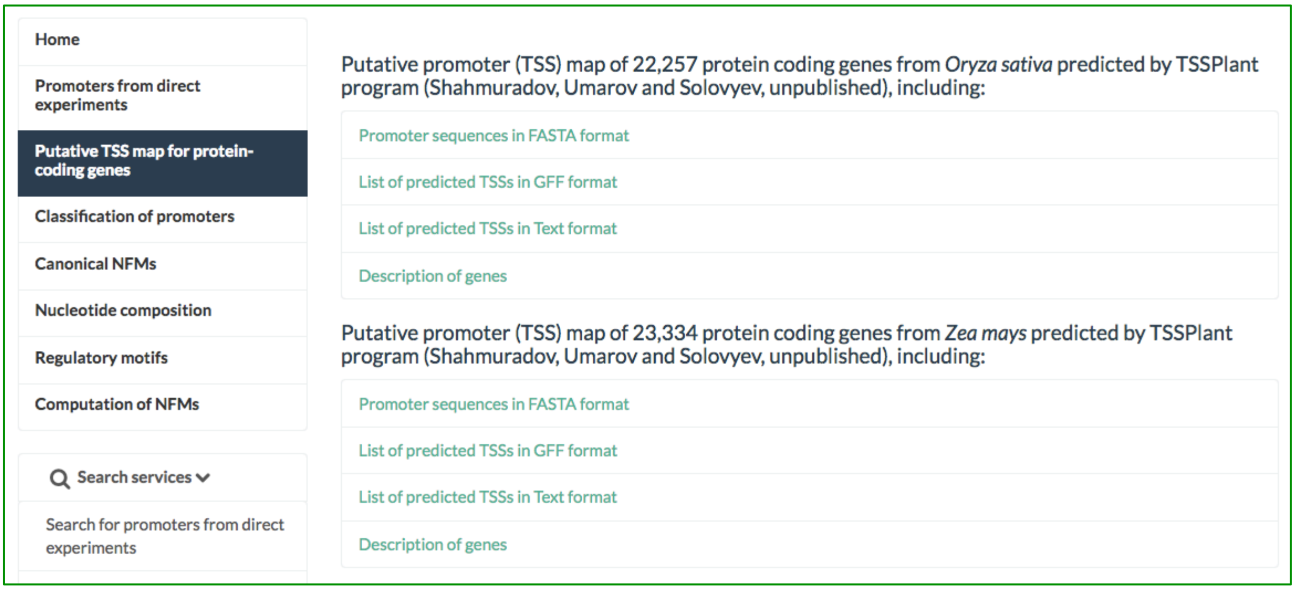
● Save File

□ Do this automatically for files like this from now on.

Click “OK”, to save GZ file with sequences.

View and download sequences of [-1000:+100] regions of protein-coding genes (+1 is an annotated gene start) and computationally predicted TSS maps for 5 genomes

In Main Menu, if an option **Putative TSS map for protein-coding genes** is chosen, the following sub-menu appears (shown here only partially):


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

Putative promoter (TSS) map of 22,257 protein coding genes from *Oryza 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

Putative promoter (TSS) map of 23,334 protein coding genes from *Zea mays* predicted by TSSPlant program (Shahmuradov, Umarov and Solovyev, unpublished), including:

Promoter sequences in FASTA format

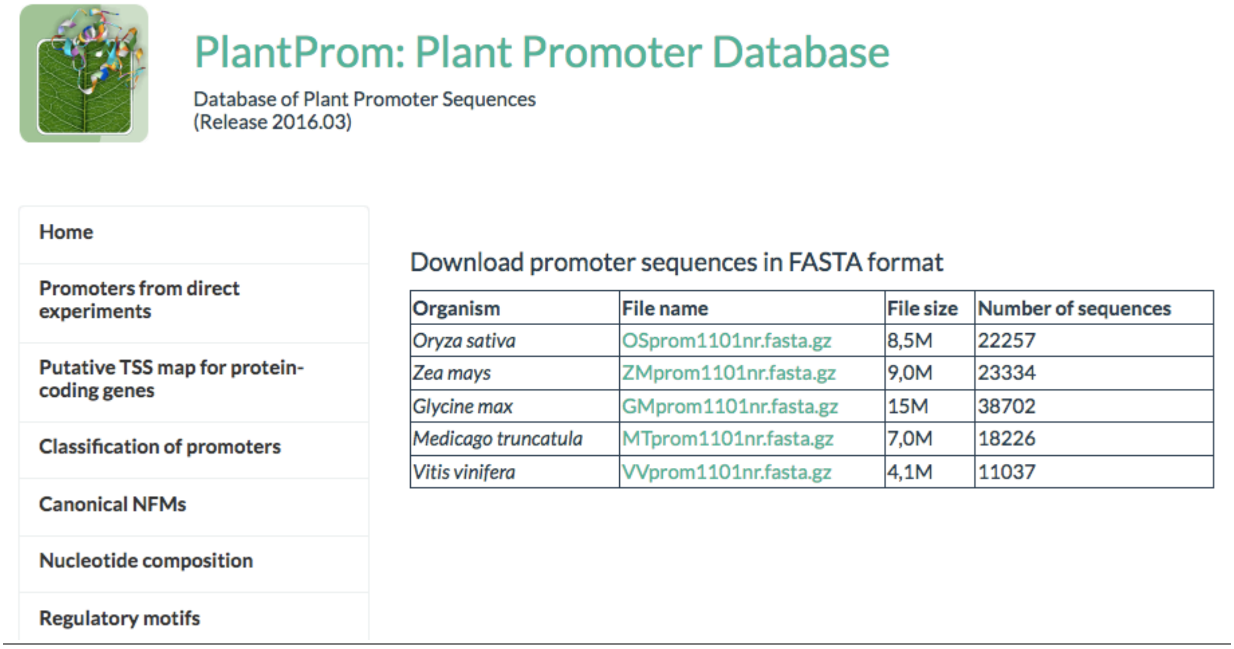
List of predicted TSSs in GFF format

List of predicted TSSs in Text format

Description of genes

In this sub-menu, for the selected species, the user can choose four options: **Promoter sequences in FASTA format**, **List of predicted TSSs in GFF format**, **List of predicted TSSs in Text format** and **Descriptions of genes**.

- If **Promoter sequences in FASTA format** is chosen, a new download page with information on the download file size and total number of records (sequences) in this file is displayed.


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

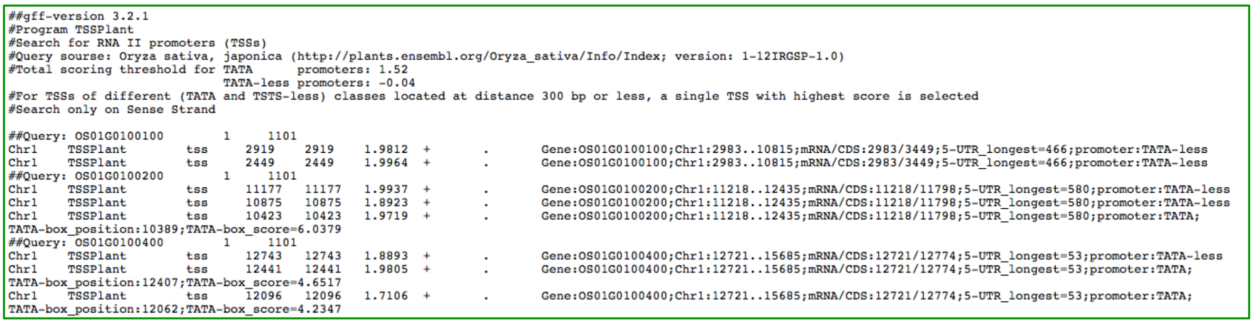
PlantProm: Plant Promoter Database

Database of Plant Promoter Sequences
(Release 2016.03)

Download promoter sequences in FASTA format

Organism	File name	File size	Number of sequences
Oryza sativa	OSprom1101nr.fasta.gz	8,5M	22257
Zea mays	ZMprom1101nr.fasta.gz	9,0M	23334
Glycine max	GMprom1101nr.fasta.gz	15M	38702
Medicago truncatula	MTprom1101nr.fasta.gz	7,0M	18226
Vitis vinifera	VVprom1101nr.fasta.gz	4,1M	11037

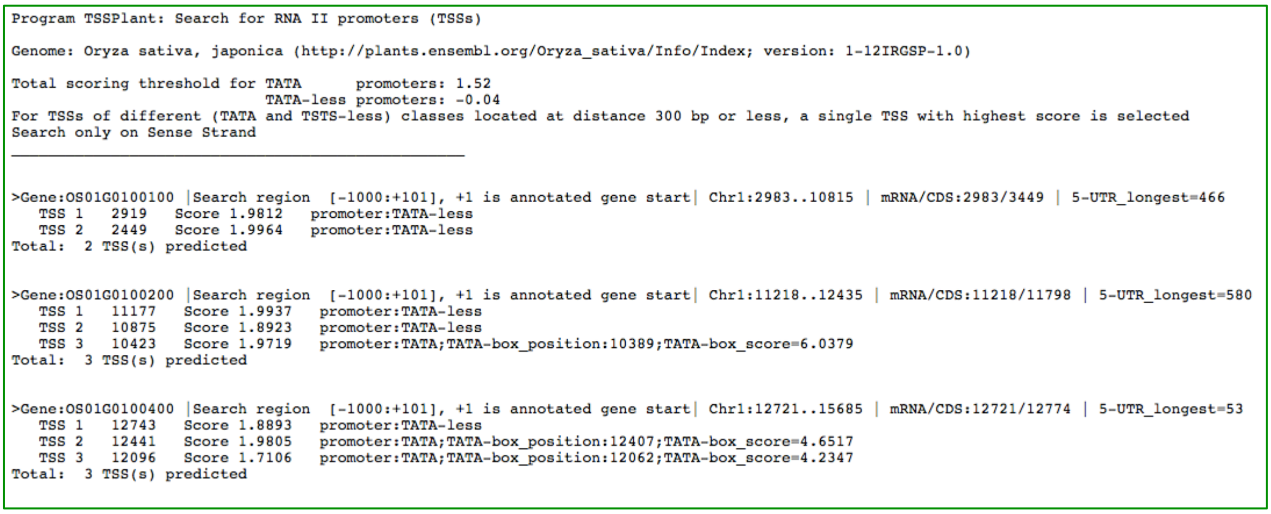
- If **List of predicted TSSs in GFF format** is chosen, a new page with data on predicted TSSs in GFF format is displayed for viewing and downloading:



```
##gff-version 3.2.1
#Program TSSPlant
#Search for RNA II promoters (TSSs)
#Query sourse: Oryza sativa, japonica (http://plants.ensembl.org/Oryza_sativa/Info/Index; version: 1-12IRGSP-1.0)
#Total scoring threshold for TATA      promoters: 1.52
                             TATA-less promoters: -0.04
#For TSSs of different (TATA and TSTS-less) classes located at distance 300 bp or less, a single TSS with highest score is selected
#Search only on Sense Strand

##Query: OS01G0100100        1     1101
Chrl    TSSPlant        tss     2919    2919    1.9812  +       .       Gene:OS01G0100100;Chr1:2983..10815;mRNA/CDS:2983/3449;5-UTR_longest=466;promoter:TATA-less
Chrl    TSSPlant        tss     2449    2449    1.9964  +       .       Gene:OS01G0100100;Chr1:2983..10815;mRNA/CDS:2983/3449;5-UTR_longest=466;promoter:TATA-less
##Query: OS01G0100200        1     1101
Chrl    TSSPlant        tss     11177   11177   1.9937  +       .       Gene:OS01G0100200;Chr1:11218..12435;mRNA/CDS:11218/11798;5-UTR_longest=580;promoter:TATA-less
Chrl    TSSPlant        tss     10875   10875   1.8923  +       .       Gene:OS01G0100200;Chr1:11218..12435;mRNA/CDS:11218/11798;5-UTR_longest=580;promoter:TATA-less
Chrl    TSSPlant        tss     10423   10423   1.9719  +       .       Gene:OS01G0100200;Chr1:11218..12435;mRNA/CDS:11218/11798;5-UTR_longest=580;promoter:TATA;
TATA-box_position:10389;TATA-box_score=6.0379
##Query: OS01G0100400        1     1101
Chrl    TSSPlant        tss     12743   12743   1.8893  +       .       Gene:OS01G0100400;Chr1:12721..15685;mRNA/CDS:12721/12774;5-UTR_longest=53;promoter:TATA-less
Chrl    TSSPlant        tss     12441   12441   1.9805  +       .       Gene:OS01G0100400;Chr1:12721..15685;mRNA/CDS:12721/12774;5-UTR_longest=53;promoter:TATA;
TATA-box_position:12407;TATA-box_score=4.6517
Chrl    TSSPlant        tss     12096   12096   1.7106  +       .       Gene:OS01G0100400;Chr1:12721..15685;mRNA/CDS:12721/12774;5-UTR_longest=53;promoter:TATA;
TATA-box_position:12062;TATA-box_score=4.2347
```

Here, TSS data for every query begins with "##Query…". Several next lines (until next query record) contain the following data: chromosome, TSS position (start and end positions are the same), Integral score for the TSS predicted, gene name, chromosome strand (+ or -), chromosome positions of gene start and end, mRNA and CDS start positions, length of the longest 5'-UTR and promoter class (TATA or TATA-less) as well as, for TATA promoters, start position and score of TATA-box.

- If **List of predicted TSSs in Text format** is clicked, a new page with data on predicted TSSs in text format is displayed:

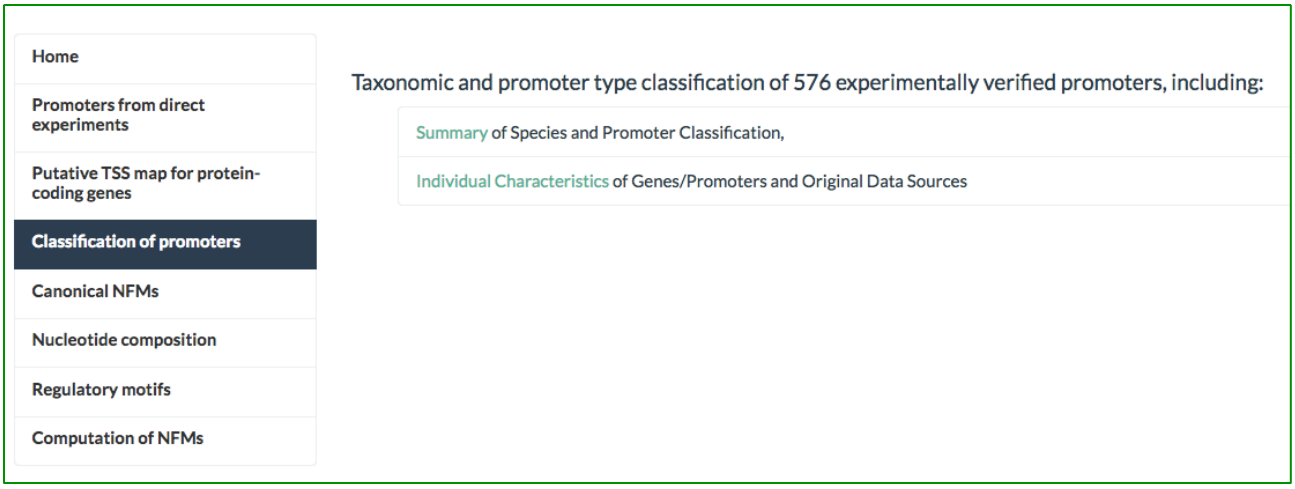


```
Program TSSPlant: Search for RNA II promoters (TSSs)

Genome: Oryza sativa, japonica (http://plants.ensembl.org/Oryza_sativa/Info/Index; version: 1-12IRGSP-1.0)

Total scoring threshold for TATA      promoters: 1.52
                            TATA-less promoters: -0.04
For TSSs of different (TATA and TSTS-less) classes located at distance 300 bp or less, a single TSS with highest score is selected
Search only on Sense Strand
________________________________________________


>Gene:OS01G0100100 |Search region  [-1000:+101], +1 is annotated gene start| Chr1:2983..10815 | mRNA/CDS:2983/3449 | 5-UTR_longest=466
   TSS 1    2919   Score 1.9812   promoter:TATA-less
   TSS 2    2449   Score 1.9964   promoter:TATA-less
Total:  2 TSS(s) predicted


>Gene:OS01G0100200 |Search region  [-1000:+101], +1 is annotated gene start| Chr1:11218..12435 | mRNA/CDS:11218/11798 | 5-UTR_longest=580
   TSS 1    11177   Score 1.9937   promoter:TATA-less
   TSS 2    10875   Score 1.8923   promoter:TATA-less
   TSS 3    10423   Score 1.9719   promoter:TATA;TATA-box_position:10389;TATA-box_score=6.0379
Total:  3 TSS(s) predicted


>Gene:OS01G0100400 |Search region  [-1000:+101], +1 is annotated gene start| Chr1:12721..15685 | mRNA/CDS:12721/12774 | 5-UTR_longest=53
   TSS 1    12743   Score 1.8893   promoter:TATA-less
   TSS 2    12441   Score 1.9805   promoter:TATA;TATA-box_position:12407;TATA-box_score=4.6517
   TSS 3    12096   Score 1.7106   promoter:TATA;TATA-box_position:12062;TATA-box_score=4.2347
Total:  3 TSS(s) predicted
```

- If **Descriptions of genes** is clicked, a new page with descriptions of genes is displayed.

View and download data on classification of 576 experimentally verified promoters by promoter class and taxonomy

On Main Menu, if an option **Classification of promoters** is chosen, the following sub-menu is displayed, consisting of two options, **Summary** and **Individual Characteristics**:

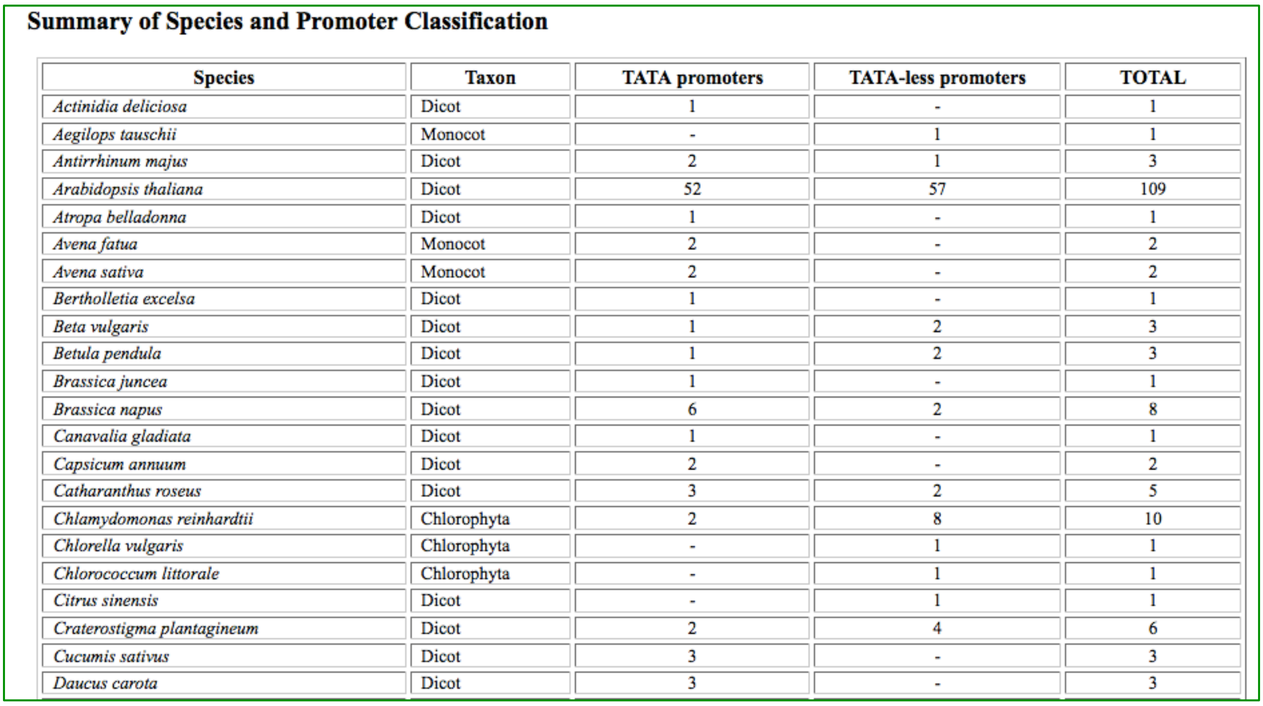

Home
Promoters from direct experiments
Putative TSS map for protein-coding genes
Classification of promoters
Canonical NFMs
Nucleotide composition
Regulatory motifs
Computation of NFMs

Taxonomic and promoter type classification of 576 experimentally verified promoters, including:

Summary of Species and Promoter Classification,

Individual Characteristics of Genes/Promoters and Original Data Sources

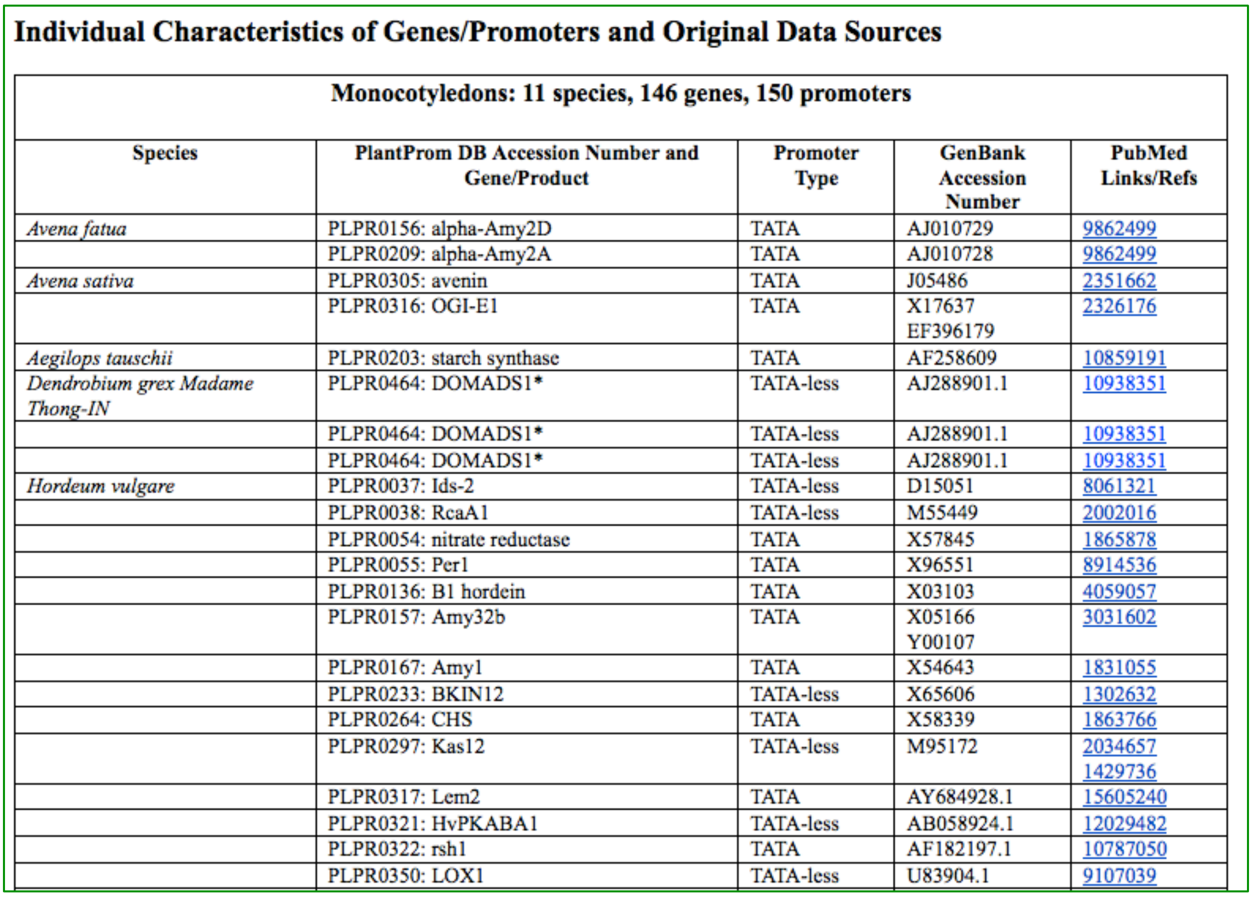
- **Summary** option displays, a new page with a list of species represented in the experimentally verified promoter set, as well as total numbers and classes of promoters form each species:



Summary of Species and Promoter Classification

Species	Taxon	TATA promoters	TATA-less promoters	TOTAL
Actinidia deliciosa	Dicot	1	-	1
Aegilops tauschii	Monocot	-	1	1
Antirrhinum majus	Dicot	2	1	3
Arabidopsis thaliana	Dicot	52	57	109
Atropa belladonna	Dicot	1	-	1
Avena fatua	Monocot	2	-	2
Avena sativa	Monocot	2	-	2
Bertholletia excelsa	Dicot	1	-	1
Beta vulgaris	Dicot	1	2	3
Betula pendula	Dicot	1	2	3
Brassica juncea	Dicot	1	-	1
Brassica napus	Dicot	6	2	8
Canavalia gladiata	Dicot	1	-	1
Capsicum annuum	Dicot	2	-	2
Catharanthus roseus	Dicot	3	2	5
Chlamydomonas reinhardtii	Chlorophyta	2	8	10
Chlorella vulgaris	Chlorophyta	-	1	1
Chlorococcum littorale	Chlorophyta	-	1	1
Citrus sinensis	Dicot	-	1	1
Craterostigma plantagineum	Dicot	2	4	6
Cucumis sativus	Dicot	3	-	3
Daucus carota	Dicot	3	-	3

- **Individual Characteristics** option loads a new page with information on genes compiled in DB, such as PlantProm DB accession number and gene/product, promoter class (type), GenBank accession number of a gene and a PubMed link to a publication that experimentally verified TSS(s) for a given gene:



Individual Characteristics of Genes/Promoters and Original Data Sources

Monocotyledons: 11 species, 146 genes, 150 promoters

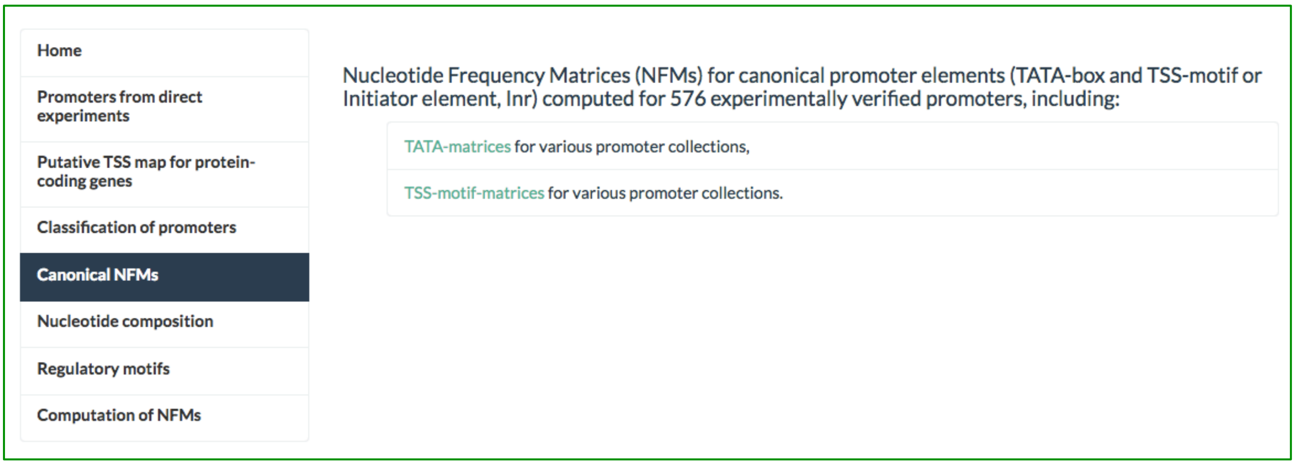
Species	PlantProm DB Accession Number and Gene/Product	Promoter Type	GenBank Accession Number	PubMed Links/Refs
Avena fatua	PLPR0156: alpha-Amy2D	TATA	AJ010729	9862499
	PLPR0209: alpha-Amy2A	TATA	AJ010728	9862499
Avena sativa	PLPR0305: avenin	TATA	J05486	2351662
	PLPR0316: OGI-E1	TATA	X17637 EF396179	2326176
Aegilops tauschii	PLPR0203: starch synthase	TATA	AF258609	10859191
Dendrobium grex Madame Thong-IN	PLPR0464: DOMADS1*	TATA-less	AJ288901.1	10938351
	PLPR0464: DOMADS1*	TATA-less	AJ288901.1	10938351
	PLPR0464: DOMADS1*	TATA-less	AJ288901.1	10938351
Hordeum vulgare	PLPR0037: Ids-2	TATA-less	D15051	8061321
	PLPR0038: RcaA1	TATA-less	M55449	2002016
	PLPR0054: nitrate reductase	TATA	X57845	1865878
	PLPR0055: Per1	TATA	X96551	8914536
	PLPR0136: B1 hordein	TATA	X03103	4059057
	PLPR0157: Amy32b	TATA	X05166 Y00107	3031602
	PLPR0167: Amy1	TATA	X54643	1831055
	PLPR0233: BKIN12	TATA-less	X65606	1302632
	PLPR0264: CHS	TATA	X58339	1863766
	PLPR0297: Kas12	TATA-less	M95172	2034657 1429736
	PLPR0317: Lem2	TATA	AY684928.1	15605240
	PLPR0321: HvPKABA1	TATA-less	AB058924.1	12029482
	PLPR0322: rsh1	TATA	AF182197.1	10787050
	PLPR0350: LOX1	TATA-less	U83904.1	9107039

Get a PubMed link for every entry of 576 experimentally verified promoters

In Main Menu, go: **Classification of promoters** ➔ **Individual Characteristics** as described above.

Retrieve and download TATA-box and INR NFMs

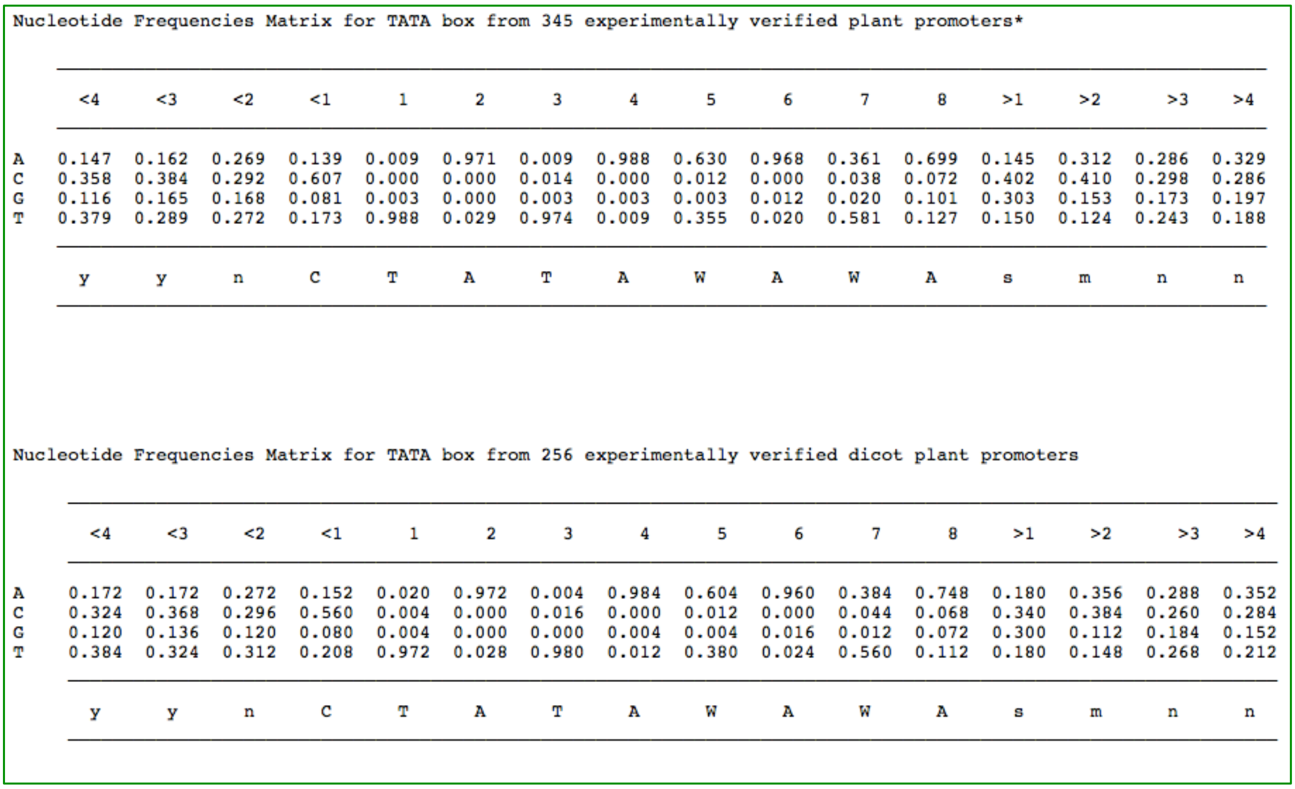
In Main Menu, an option **Canonical NFMs** displays the following sub-menu with two options, **TATA-matrices** and **TSS-motif-matrices**:


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

Nucleotide Frequency Matrices (NFMs) for canonical promoter elements (TATA-box and TSS-motif or Initiator element, Inr) computed for 576 experimentally verified promoters, including:

- TATA-matrices for various promoter collections,
- TSS-motif-matrices for various promoter collections.

- **TATA-matrices** option loads a page with TATA-matrices for various promoter collections (here shown only partially):



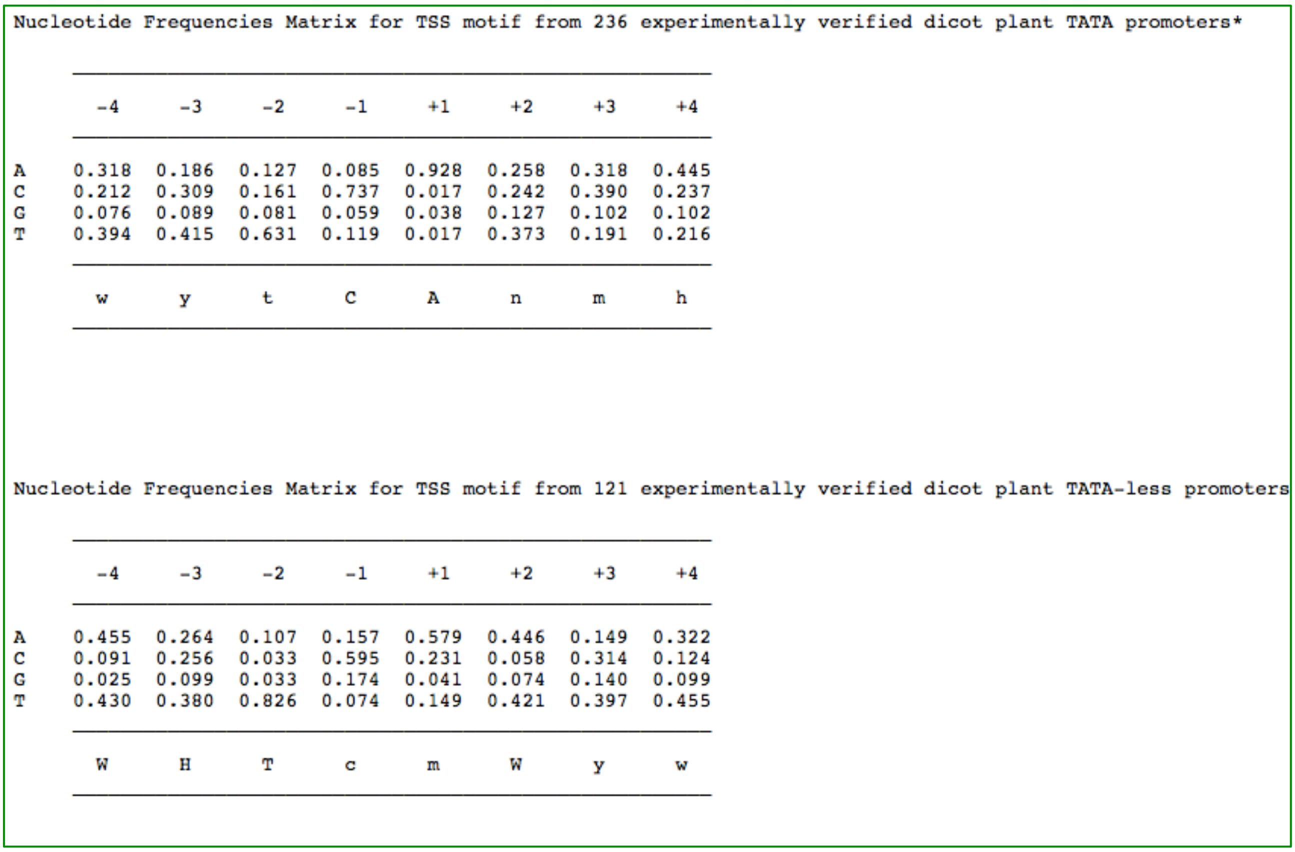
Nucleotide Frequencies Matrix for TATA box from 345 experimentally verified plant promoters*

	<4	<3	<2	<1	1	2	3	4	5	6	7	8	>1	>2	>3	>4
A	0.147	0.162	0.269	0.139	0.009	0.971	0.009	0.988	0.630	0.968	0.361	0.699	0.145	0.312	0.286	0.329
C	0.358	0.384	0.292	0.607	0.000	0.000	0.014	0.000	0.012	0.000	0.038	0.072	0.402	0.410	0.298	0.286
G	0.116	0.165	0.168	0.081	0.003	0.000	0.003	0.003	0.003	0.012	0.020	0.101	0.303	0.153	0.173	0.197
T	0.379	0.289	0.272	0.173	0.988	0.029	0.974	0.009	0.355	0.020	0.581	0.127	0.150	0.124	0.243	0.188
	y	y	n	C	T	A	T	A	W	A	W	A	s	m	n	n

Nucleotide Frequencies Matrix for TATA box from 256 experimentally verified dicot plant promoters

	<4	<3	<2	<1	1	2	3	4	5	6	7	8	>1	>2	>3	>4
A	0.172	0.172	0.272	0.152	0.020	0.972	0.004	0.984	0.604	0.960	0.384	0.748	0.180	0.356	0.288	0.352
C	0.324	0.368	0.296	0.560	0.004	0.000	0.016	0.000	0.012	0.000	0.044	0.068	0.340	0.384	0.260	0.284
G	0.120	0.136	0.120	0.080	0.004	0.000	0.000	0.004	0.004	0.016	0.012	0.072	0.300	0.112	0.184	0.152
T	0.384	0.324	0.312	0.208	0.972	0.028	0.980	0.012	0.380	0.024	0.560	0.112	0.180	0.148	0.268	0.212
	y	y	n	C	T	A	T	A	W	A	W	A	s	m	n	n

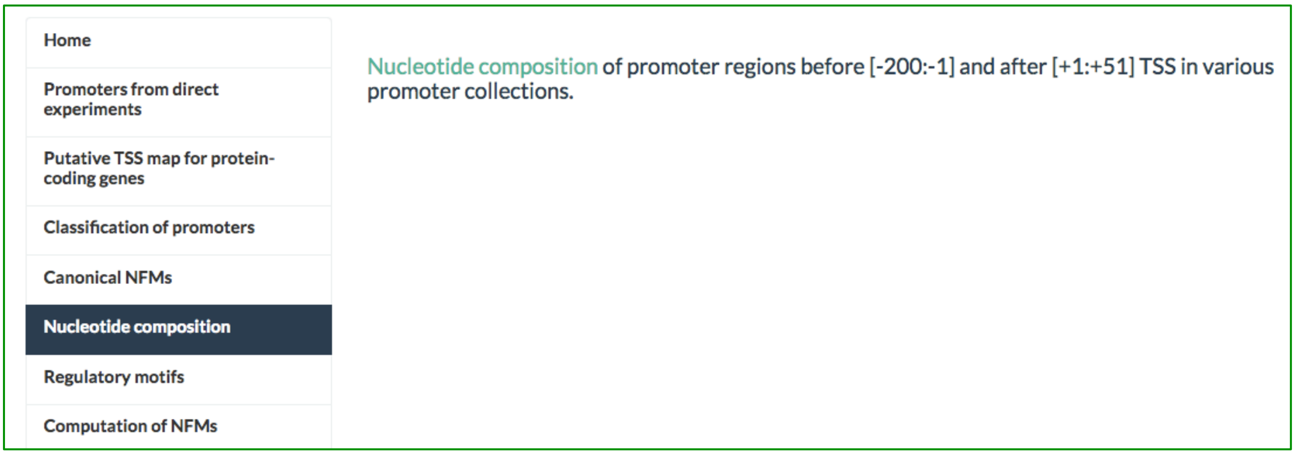
- **TSS-motif-matrices** option loads a page with TSS-motif matrices for various promoter collections (shown here partially):



```
Nucleotide Frequencies Matrix for TSS motif from 236 experimentally verified dicot plant TATA promoters*

     ______________________________________________________
       -4     -3     -2     -1     +1     +2     +3     +4
     ______________________________________________________

A    0.318  0.186  0.127  0.085  0.928  0.258  0.318  0.445
C    0.212  0.309  0.161  0.737  0.017  0.242  0.390  0.237
G    0.076  0.089  0.081  0.059  0.038  0.127  0.102  0.102
T    0.394  0.415  0.631  0.119  0.017  0.373  0.191  0.216
     ______________________________________________________
      w      y      t      C      A      n      m      h
     ______________________________________________________




Nucleotide Frequencies Matrix for TSS motif from 121 experimentally verified dicot plant TATA-less promoters

     ______________________________________________________
       -4     -3     -2     -1     +1     +2     +3     +4
     ______________________________________________________

A    0.455  0.264  0.107  0.157  0.579  0.446  0.149  0.322
C    0.091  0.256  0.033  0.595  0.231  0.058  0.314  0.124
G    0.025  0.099  0.033  0.174  0.041  0.074  0.140  0.099
T    0.430  0.380  0.826  0.074  0.149  0.421  0.397  0.455
     ______________________________________________________
      W      H      T      c      m      W      y      w
     ______________________________________________________
```

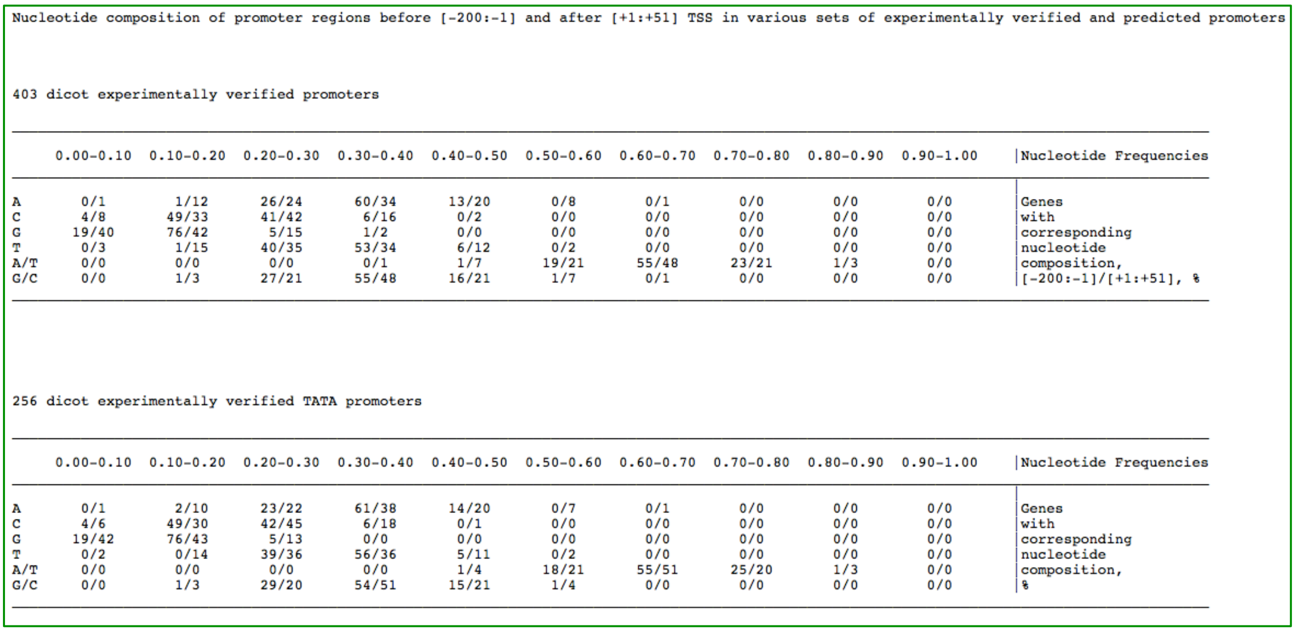
Nucleotide composition of promoter regions

In Main Menu, an option **Nucleotide composition** calls the following sub-menu with a single option, **Nucleotide composition**:


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

Nucleotide composition of promoter regions before [-200:-1] and after [+1:+51] TSS in various promoter collections.

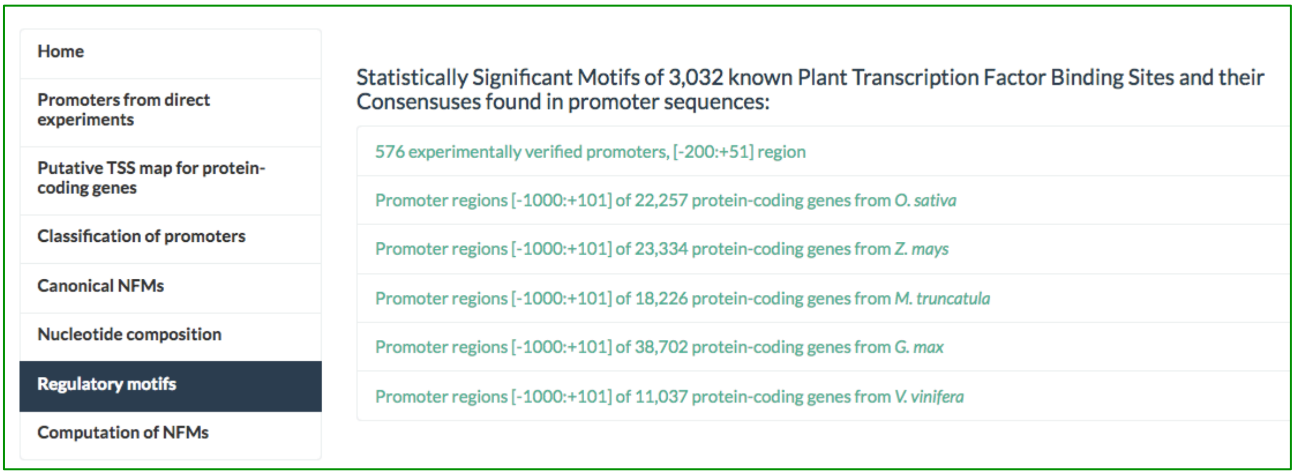
which in turn, if clicked, loads a page with nucleotide composition of promoter regions before [-200:-1], and after [+1:+51] TSSs, for various promoter sets



```
Nucleotide composition of promoter regions before [-200:-1] and after [+1:+51] TSS in various sets of experimentally verified and predicted promoters


403 dicot experimentally verified promoters

      0.00-0.10  0.10-0.20  0.20-0.30  0.30-0.40  0.40-0.50  0.50-0.60  0.60-0.70  0.70-0.80  0.80-0.90  0.90-1.00     |Nucleotide Frequencies

A        0/1       1/12      26/24      60/34      13/20       0/8        0/1        0/0        0/0        0/0       |Genes
C        4/8      49/33      41/42       6/16       0/2        0/0        0/0        0/0        0/0        0/0       |with
G       19/40     76/42       5/15       1/2        0/0        0/0        0/0        0/0        0/0        0/0       |corresponding
T        0/3       1/15      40/35      53/34       6/12       0/2        0/0        0/0        0/0        0/0       |nucleotide
A/T      0/0       0/0        0/0        0/1        1/7       19/21      55/48      23/21       1/3        0/0       |composition,
G/C      0/0       1/3       27/21      55/48      16/21       1/7        0/1        0/0        0/0        0/0       |[-200:-1]/[+1:+51], %



256 dicot experimentally verified TATA promoters

      0.00-0.10  0.10-0.20  0.20-0.30  0.30-0.40  0.40-0.50  0.50-0.60  0.60-0.70  0.70-0.80  0.80-0.90  0.90-1.00     |Nucleotide Frequencies

A        0/1       2/10      23/22      61/38      14/20       0/7        0/1        0/0        0/0        0/0       |Genes
C        4/6      49/30      42/45       6/18       0/1        0/0        0/0        0/0        0/0        0/0       |with
G       19/42     76/43       5/13       0/0        0/0        0/0        0/0        0/0        0/0        0/0       |corresponding
T        0/2       0/14      39/36      56/36       5/11       0/2        0/0        0/0        0/0        0/0       |nucleotide
A/T      0/0       0/0        0/0        0/0        1/4       18/21      55/51      25/20       1/3        0/0       |composition,
G/C      0/0       1/3       29/20      54/51      15/21       1/4        0/0        0/0        0/0        0/0       |%
```

Retrieve and download putative TFBS content of promoter sequences

In Main Menu, an option **Regulatory motifs** displays the following sub-menu with six options for 576 experimentally verified promoters and promoter regions for five species, as *O. sativa*, *Z. mays*, *M. truncatula*, *G. max* and *V. vinifera*:


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

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

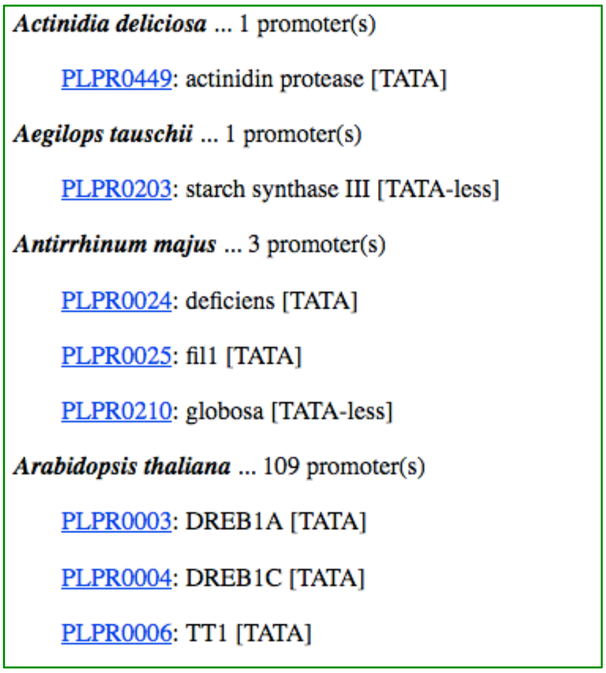
Promoter regions [-1000:+101] of 23,334 protein-coding genes from *Z. mays*

Promoter regions [-1000:+101] of 18,226 protein-coding genes from *M. truncatula*

Promoter regions [-1000:+101] of 38,702 protein-coding genes from *G. max*

Promoter regions [-1000:+101] of 11,037 protein-coding genes from *V. vinifera*

- An option **576 experimentally verified promoters, [-200:+51] regions**, shows a list of 576 genes:



Actinidia deliciosa ... 1 promoter(s)

PLPR0449: actinidin protease [TATA]

Aegilops tauschii ... 1 promoter(s)

PLPR0203: starch synthase III [TATA-less]

Antirrhinum majus ... 3 promoter(s)

PLPR0024: deficiens [TATA]

PLPR0025: fil1 [TATA]

PLPR0210: globosa [TATA-less]

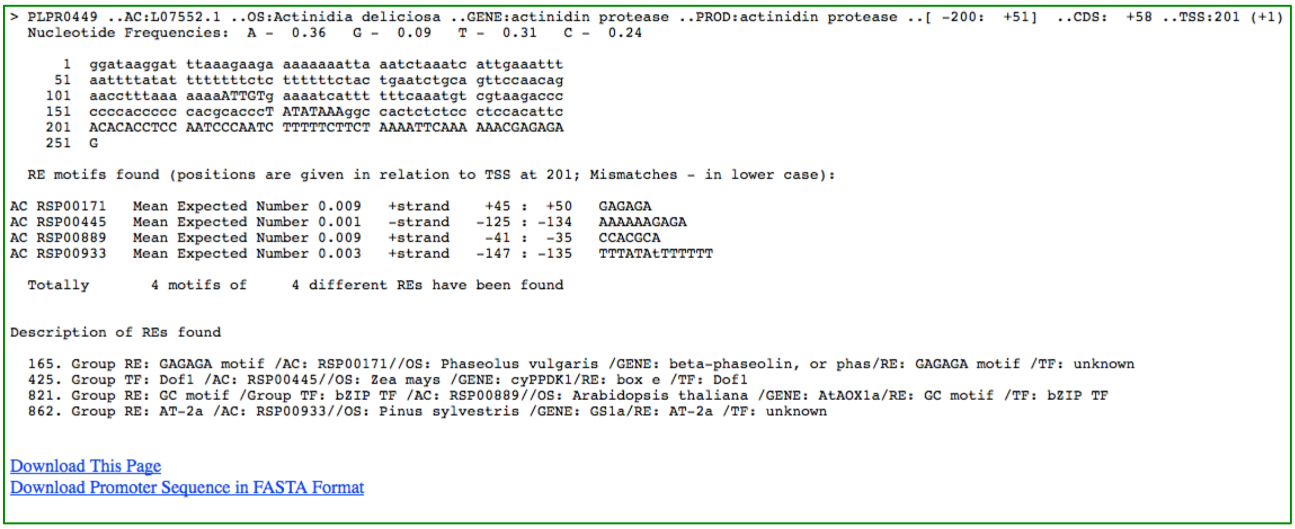
Arabidopsis thaliana ... 109 promoter(s)

PLPR0003: DREB1A [TATA]

PLPR0004: DREB1C [TATA]

PLPR0006: TT1 [TATA]

Here, by clicking on PlantProm DB ID of a promoter (e.g. **PLPR0449** the data on statistically non-random motifs of 3,032 known plant transcription factor binding sites (TFBSs), predicted by Nsite program (Shahmuradov and Solovyev, Bioinformatics, 2015, 21:3544; see also **Related Links** option in Main Menu), can be viewed and downloaded:

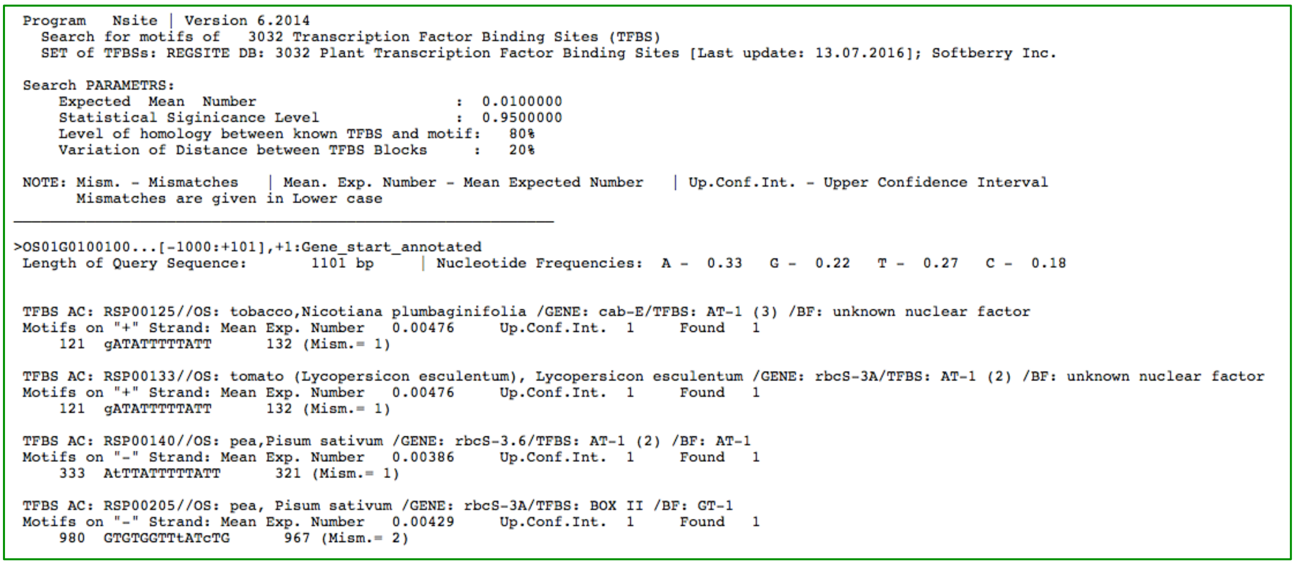


```
> PLPR0449 ..AC:L07552.1 ..OS:Actinidia deliciosa ..GENE:actinidin protease ..PROD:actinidin protease ..[ -200:  +51]  ..CDS:  +58 ..TSS:201 (+1)
  Nucleotide Frequencies:  A -  0.36   G -  0.09   T -  0.31   C -  0.24

       1  ggataaggat ttaaagaaga aaaaaaatta aatctaaatc attgaaattt
      51  aattttatat tttttttctc ttttttctac tgaatctgca gttccaacag
     101  aacctttaaa aaaaATTGTg aaaatcattt tttcaaatgt cgtaagaccc
     151  ccccaccccc cacgcacccT ATATAAAggc cactctctcc ctccacattc
     201  ACACACCTCC AATCCCAATC TTTTTCTTCT AAAATTCAAA AAACGAGAGA
     251  G

  RE motifs found (positions are given in relation to TSS at 201; Mismatches - in lower case):

AC RSP00171   Mean Expected Number 0.009   +strand    +45 :  +50   GAGAGA
AC RSP00445   Mean Expected Number 0.001   -strand   -125 : -134   AAAAAAGAGA
AC RSP00889   Mean Expected Number 0.009   +strand    -41 :  -35   CCACGCA
AC RSP00933   Mean Expected Number 0.003   +strand   -147 : -135   TTTATAtTTTTTT

  Totally      4 motifs of     4 different REs have been found


Description of REs found

  165. Group RE: GAGAGA motif /AC: RSP00171//OS: Phaseolus vulgaris /GENE: beta-phaseolin, or phas/RE: GAGAGA motif /TF: unknown
  425. Group TF: Dof1 /AC: RSP00445//OS: Zea mays /GENE: cyPPDK1/RE: box e /TF: Dof1
  821. Group RE: GC motif /Group TF: bZIP TF /AC: RSP00889//OS: Arabidopsis thaliana /GENE: AtAOX1a/RE: GC motif /TF: bZIP TF
  862. Group RE: AT-2a /AC: RSP00933//OS: Pinus sylvestris /GENE: GS1a/RE: AT-2a /TF: unknown
```

Download This Page
Download Promoter Sequence in FASTA Format

This page contains two options, **Download This Page** and **Download Promoter Sequence in FASTA format**.

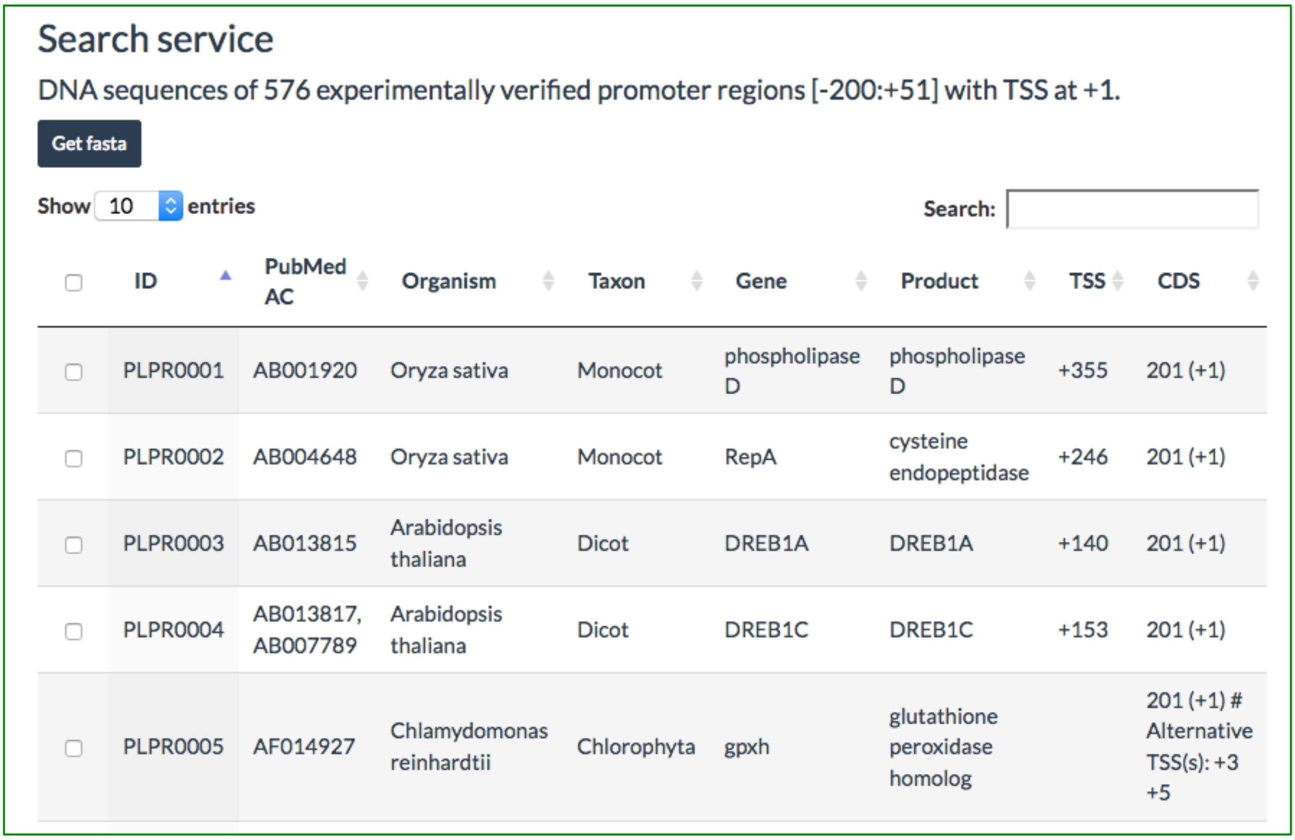
- If one of the next five options of the sub-menu is chosen (e.g. **Promoter regions [-1000: +101] of 22,257 protein-coding genes…** (from *O. sativa* or another species), statistically non-random motifs of known TFBSs found in every gene of that species are displayed:



```
 Program   Nsite | Version 6.2014
   Search for motifs of   3032 Transcription Factor Binding Sites (TFBS)
   SET of TFBSs: REGSITE DB: 3032 Plant Transcription Factor Binding Sites [Last update: 13.07.2016]; Softberry Inc.

 Search PARAMETRS:
     Expected  Mean  Number                     :  0.0100000
     Statistical Siginicance Level              :  0.9500000
     Level of homology between known TFBS and motif:   80%
     Variation of Distance between TFBS Blocks     :   20%

 NOTE: Misn. - Mismatches    | Mean. Exp. Number - Mean Expected Number   | Up.Conf.Int. - Upper Confidence Interval
       Mismatches are given in Lower case
____________________________________________________________

>OS01G0100100...[-1000:+101],+1:Gene_start_annotated
 Length of Query Sequence:       1101 bp     | Nucleotide Frequencies:  A -  0.33   G -  0.22   T -  0.27   C -  0.18


 TFBS AC: RSP00125//OS: tobacco,Nicotiana plumbaginifolia /GENE: cab-E/TFBS: AT-1 (3) /BF: unknown nuclear factor
 Motifs on "+" Strand: Mean Exp. Number   0.00476     Up.Conf.Int.  1     Found   1
     121  gATATTTTTATT      132 (Mism.= 1)

 TFBS AC: RSP00133//OS: tomato (Lycopersicon esculentum), Lycopersicon esculentum /GENE: rbcS-3A/TFBS: AT-1 (2) /BF: unknown nuclear factor
 Motifs on "+" Strand: Mean Exp. Number   0.00476     Up.Conf.Int.  1     Found   1
     121  gATATTTTTATT      132 (Mism.= 1)

 TFBS AC: RSP00140//OS: pea,Pisum sativum /GENE: rbcS-3.6/TFBS: AT-1 (2) /BF: AT-1
 Motifs on "-" Strand: Mean Exp. Number   0.00386     Up.Conf.Int.  1     Found   1
     333  AtTTATTTTTATT      321 (Mism.= 1)

 TFBS AC: RSP00205//OS: pea, Pisum sativum /GENE: rbcS-3A/TFBS: BOX II /BF: GT-1
 Motifs on "-" Strand: Mean Exp. Number   0.00429     Up.Conf.Int.  1     Found   1
     980  GTGTGGTTtATcTG      967 (Mism.= 2)
```

Search for experimentally verified promoters by PlantProm DB ID

In **Search services** of Main Menu, if an option **Search for promoters from direct experiments** is chosen, the following page is appears:



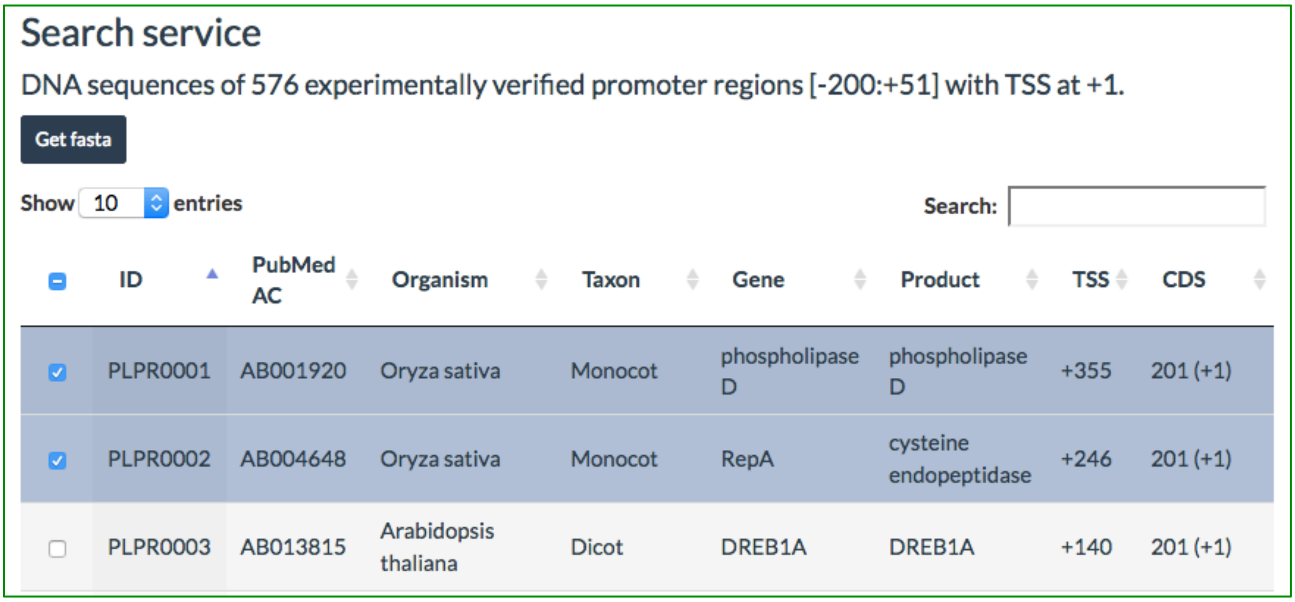
Search service

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

□	ID	PubMed AC	Organism	Taxon	Gene	Product	TSS	CDS
□	PLPR0001	AB001920	Oryza sativa	Monocot	phospholipase D	phospholipase D	+355	201 (+1)
□	PLPR0002	AB004648	Oryza sativa	Monocot	RepA	cysteine endopeptidase	+246	201 (+1)
□	PLPR0003	AB013815	Arabidopsis thaliana	Dicot	DREB1A	DREB1A	+140	201 (+1)
□	PLPR0004	AB013817, AB007789	Arabidopsis thaliana	Dicot	DREB1C	DREB1C	+153	201 (+1)
□	PLPR0005	AF014927	Chlamydomonas reinhardtii	Chlorophyta	gpxh	glutathione peroxidase homolog		201 (+1) # Alternative TSS(s): +3 +5

Here, one or several promoters can be selected by (1) checking corresponding boxes to the left or or (2) performing search by a keyword, e.g. **PLPR057** (see pictures below). The following search options are applied: “ID” – promoter ID in DB; “Organism” – name of species (e.g.

Oryza sativa; “Taxon” – taxonomic group (e.g. Monocot); “Gene” – full name of a gene or a phrase included by the gene name; “Product” – full name of a gene product or a phrase included by gene product name; for the full list of species and taxonomic groups see: http://www.softberry.com/data/plantprom/Links/Taxon_Table_1.htm.



Search service

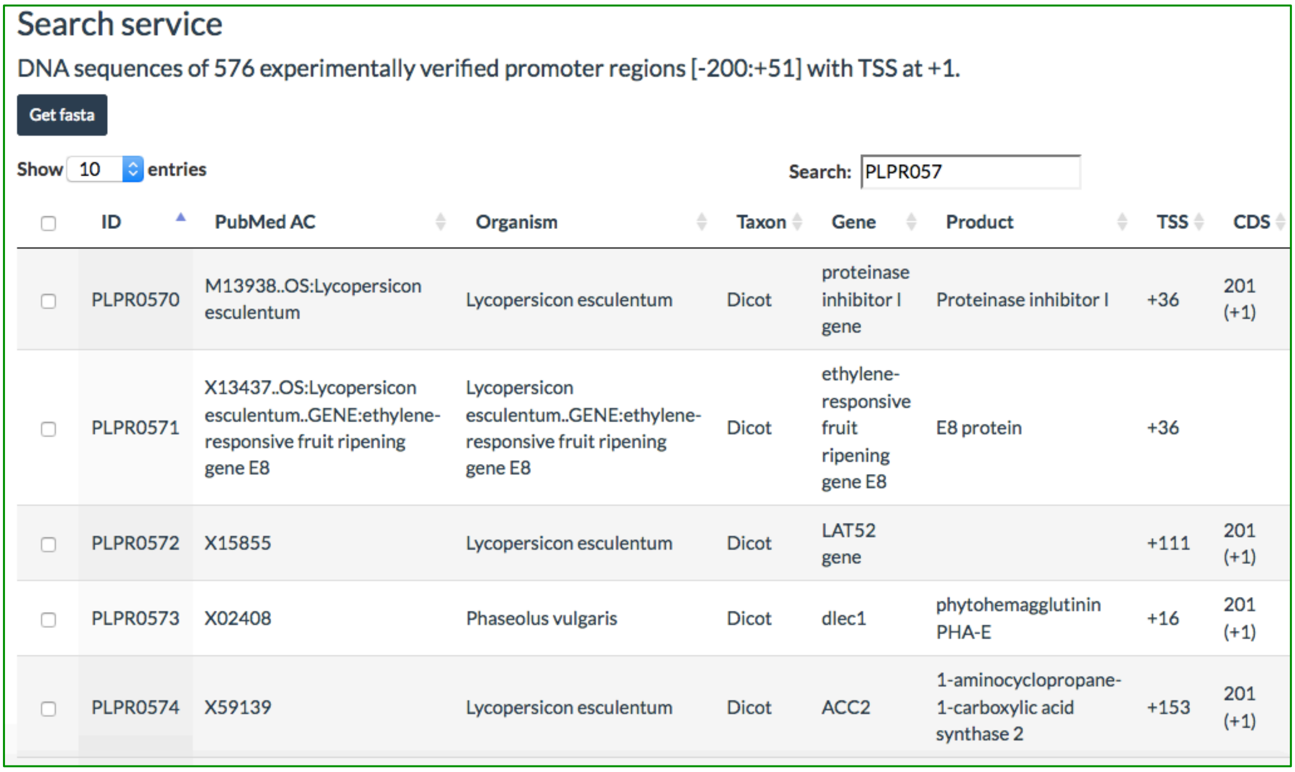
DNA sequences of 576 experimentally verified promoter regions [-200:+51] with TSS at +1.

Get fasta

Show 10 entries

Search:

☐	ID	PubMed AC	Organism	Taxon	Gene	Product	TSS	CDS
☑	PLPR0001	AB001920	Oryza sativa	Monocot	phospholipase D	phospholipase D	+355	201 (+1)
☑	PLPR0002	AB004648	Oryza sativa	Monocot	RepA	cysteine endopeptidase	+246	201 (+1)
☐	PLPR0003	AB013815	Arabidopsis thaliana	Dicot	DREB1A	DREB1A	+140	201 (+1)



Search service

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

Get fasta

Show 10 entries

Search: PLPR057

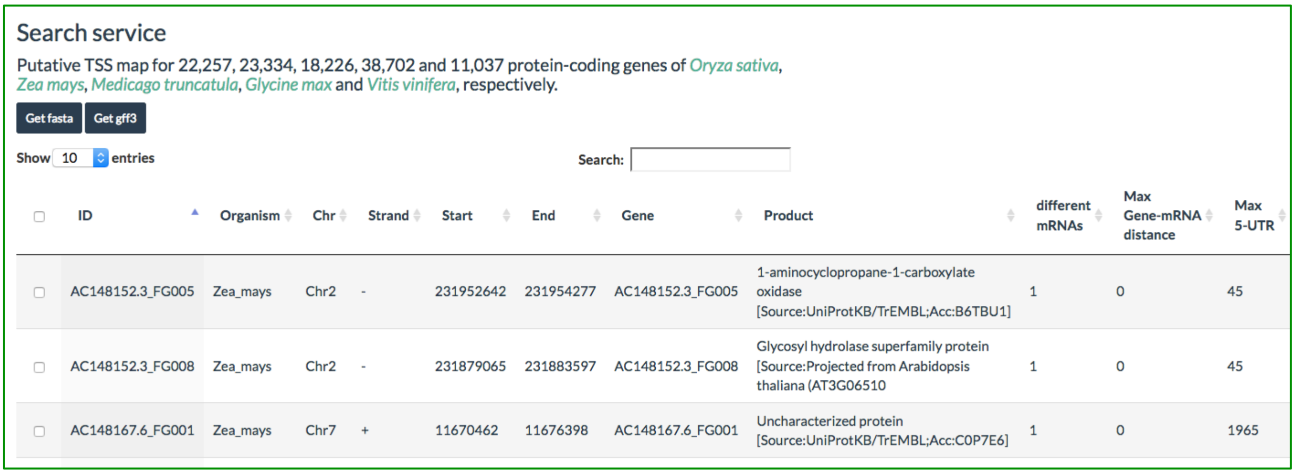
☐	ID	PubMed AC	Organism	Taxon	Gene	Product	TSS	CDS
☐	PLPR0570	M13938..OS:Lycopersicon esculentum	Lycopersicon esculentum	Dicot	proteinase inhibitor I gene	Proteinase inhibitor I	+36	201 (+1)
☐	PLPR0571	X13437..OS:Lycopersicon esculentum..GENE:ethylene-responsive fruit ripening gene E8	Lycopersicon esculentum..GENE:ethylene-responsive fruit ripening gene E8	Dicot	ethylene-responsive fruit ripening gene E8	E8 protein	+36	
☐	PLPR0572	X15855	Lycopersicon esculentum	Dicot	LAT52 gene		+111	201 (+1)
☐	PLPR0573	X02408	Phaseolus vulgaris	Dicot	dlec1	phytohemagglutinin PHA-E	+16	201 (+1)
☐	PLPR0574	X59139	Lycopersicon esculentum	Dicot	ACC2	1-aminocyclopropane-1-carboxylic acid synthase 2	+153	201 (+1)

Afterwards, if Get fasta button is clicked, a page with FASTA sequences of selected promoters appears.

Gene list can be sorted by by GenBank accession number, organism name, gene name and gene product.

Search for putative TSS map for 22,257, 23,334, 18,226, 38,702 and 11,037 protein-coding genes of five species

In **Search services** option of Main Menu, click on **Search for putative TSS map for protein-coding genes**, and the following page is displayed:



Search service

Putative TSS map for 22,257, 23,334, 18,226, 38,702 and 11,037 protein-coding genes of *Oryza sativa*, *Zea mays*, *Medicago truncatula*, *Glycine max* and *Vitis vinifera*, respectively.

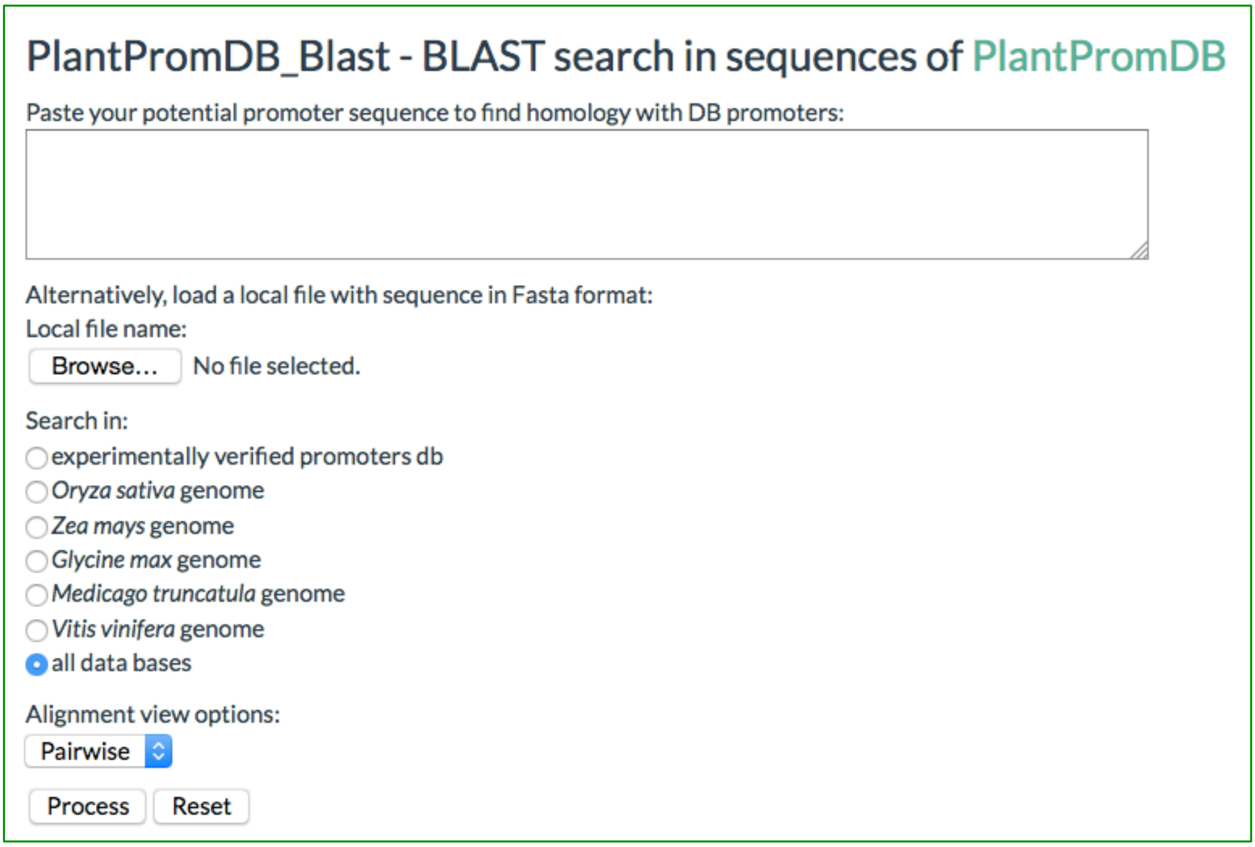
□	ID	Organism	Chr	Strand	Start	End	Gene	Product	different mRNAs	Max Gene-mRNA distance	Max 5-UTR
□	AC148152.3_FG005	Zea_mays	Chr2	-	231952642	231954277	AC148152.3_FG005	1-aminocyclopropane-1-carboxylate oxidase [Source:UniProtKB/TrEMBL;Acc:B6TBU1]	1	0	45
□	AC148152.3_FG008	Zea_mays	Chr2	-	231879065	231883597	AC148152.3_FG008	Glycosyl hydrolase superfamily protein [Source:Projected from Arabidopsis thaliana (AT3G06510	1	0	45
□	AC148167.6_FG001	Zea_mays	Chr7	+	11670462	11676398	AC148167.6_FG001	Uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:C0P7E6]	1	0	1965

Here, the following search options are applied: “ID” – promoter ID in the corresponding Ensembl genome annotation; “Organism” – one of five species (*Oryza sativa*, *Zea mays*, *Medicago truncatula*, *Glycine max* and *Vitis vinifera*); “Chr” – chromosome number (e.g. Chr 1); “Gene” – gene name accordingly to the Ensembl genome annotation; “Product” – full name of a gene product or a phrase included by gene product name; “different mRNAs” – number of alternative mRNAs from the corresponding Ensembl genome annotation. The selected promoters can be viewed and downloaded in two popular formats: FASTA (click on Get fasta and gff (click on Get gff3.

Moreover, the gene list can be sorted by gene ID, organism name, chromosome number, DNA strand, gene start position on chromosome, gene name, gene product and number of different mRNAs.

Perform BLAST comparison of user-given query sequence with promoter sequences collected in DB

In **Search services** option of Main Menu, if **BLAST search** option is chosen, the following page is displayed:



PlantPromDB_Blast - BLAST search in sequences of PlantPromDB

Paste your potential promoter sequence to find homology with DB promoters:

Alternatively, load a local file with sequence in Fasta format:
Local file name:

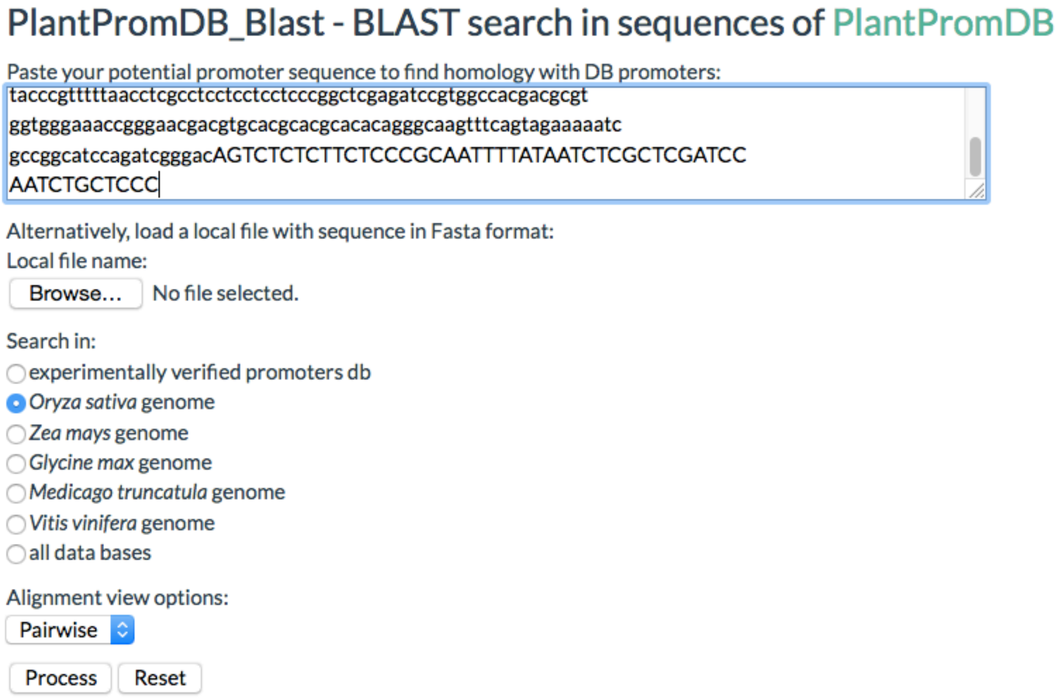
No file selected.

Search in:
○ experimentally verified promoters db
○ *Oryza sativa* genome
○ *Zea mays* genome
○ *Glycine max* genome
○ *Medicago truncatula* genome
○ *Vitis vinifera* genome
● all data bases

Alignment view options:
Pairwise

To perform BLAST search, (1) Paste a query sequence in FASTA format or browse and select a file from the corresponding folder; (2) Choose a promoter set from the list given below; (3) Choose the alignment option (**Pairwise** or **Tabular**); and finally click **Process** button.

For example:



PlantPromDB_Blast - BLAST search in sequences of PlantPromDB

Paste your potential promoter sequence to find homology with DB promoters:

```
tacccgtttttaacctcgcctcctcctcctcccggctcgagatccgtggccacgacgcgt
ggtgggaaaccgggaacgacgtgcacgcacgcacacagggcaagtttcagtagaaaaatc
gccggcatccagatcgggacAGTCTCTCTTCTCCCGCAATTTTATAATCTCGCTCGATCC
AATCTGCTCCC
```

Alternatively, load a local file with sequence in Fasta format:
Local file name:

No file selected.

Search in:
○ experimentally verified promoters db
● *Oryza sativa* genome
○ *Zea mays* genome
○ *Glycine max* genome
○ *Medicago truncatula* genome
○ *Vitis vinifera* genome
○ all data bases

Alignment view options:
Pairwise

Description of the header of FASTA files with promoter sequences in Module "Promoters from direct experiments"

The header of FASTA files contains the following information:

- **PLPRXXXX** : promoter ID in the DB;
- **AC**: GenBank accession number of a promoter;
- **OS**: name of organism/species;
- **GENE**: name of a gene;
- **PROD**: gene product;
- **[-200:+51]**: proximal promoter region including 200 bp upstream of the experimentally identified TSS (position +1) and 51 bp of the transcribed region (upper case letters);
- **Taxon**: name of the taxonomic group (Dicot, Monocot, etc.);
- **Promoter**: a class of promoter (TATA or TATA-less).

Description of the header of FASTA files for promoter sequences in Module "Putative TSS map for protein-coding genes"

The header of FASTA files contains the following information:

- **OS**: name of organism/species;
- **Chr**: the chromosome number;
- **(+) or (-)**: DNA strand of gene location;
- **xxxxxxxxx..xxxxxxxxx**: the annotated start and end positions of a gene on chromosome;
- **Gene**: name of a gene;
- **mRNA/CDS**: The annotated start position(s) of mRNA and corresponding coding sequence (CDS)*;
- **Product:** gene product;
- **different mRNAs**: number of alternative mRNAs annotated;
- **Max 5-UTR**: length of the longest 5'-untrslated region (UTR) of mRNA annotated;
- **[-1000:+101]**: promoter region including 1000 bp upstream of the annotated gene start (position +1) and 101 bp of the transcribed region.

* If two or more different mRNAs are annotated, all Gene and mRNA pairs separated by coma are given.

Short description of approaches and tools applied for computation of nucleotide frequency matrices for various promoter elements, search for plant transcription factor binding sites and prediction of putative TSSs

To get unrelated set of promoters, a pairwise comparison of a region [-50:+1] of 586 plant promoters (including 305 entries from the first release of DB) has been performed and one of the couple of promoters showing more than 90% homology has been excluded from the initial collection. As a result, 10 promoters were excluded from the initial set of the collected promoter sequences.

In simple implementation of Expectation Maximization (EM) algorithm (Cardon, Stormo, 1992) we considered the sequence of motif $X=(x_1,x_2,...,x_l)$, where l is the motif length. If $P^i(x_j)$ is the empiric frequency of the nucleotide x_j in position i (computed on previous iteration), then the weight of this motif is computed as
$W(X) = \log \prod P^i(x_j)/0.25$

Using the EM procedure for 10 iterations, the initial collection of 576 unrelated promoters was divided into the 2 classes: 345TATA and 231 TATA-less unrelated promoters. In calculations of TATA matrices the allowed variation of a distance between the right boundary of the TATA-core box and the TSS was 18-40 bp and only **TATAWAWA**-core was used for calculating the weight. As an initial TATA-box matrix, the TATA-matrix computed for 171 plant promoters from the first release of PlantProm DB (Shahmuradov et al., 2003) was used.

The TSS-motif matrix of 5 bp in length has been computed, where the 3rd nucleotide was the annotated (anTSS). No strong consensus was revealed. When the EM approach was used to analyze all possible penta-nucleotides with an assumed TSS (asTSS) location in the range [anTSS-2:anTSS+2], it was observed that the composition of asTSS-motifs is different in dicot and monocot plants, as well as for TATA and TATA-less promoters.

Search for statistically significant motifs of 1577 known plant transcription regulatory elements was performed by Nsite program (Shahmuradov, Solovyev, 2015; http://linux1.softberry.com/berry.phtml).

Search for putative TSSs in genomic sequences from was performed by TSSPlant program (Shahmuradov, Umarov and Solovyev, submitted to Nucl Acid Res).

REFERENCES

Cardon L and Stormo G (1992) Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. J. Mol. Biol., 5, 159–170 (PMID: 1731067).

Shahmuradov IA, Gammerman AJ, Hancock JM, Bramley PM, Solovyev VV (2003) PlantProm: a database of plant promoter sequences. Nucleic Acids Res., 31: 114-117 (PMID: 12519961).

Shahmuradov IA, Solovyev VV (2015) Nsite, NsiteH and NsiteM computer tools for studying transcription regulatory elements. Bioinformatics, 31: 3544-3545 (PMID: 26142184).