**SUPPLEMENTARY MATERIALS**

**A new tool TSSPlant for prediction of plant Pol II promoters**
I.A.Shahmuradov, R.Kh.Umarov, V.V.Solovyev

Table S1. Promoter sets used in training and testing procedures

| Source    | Species          | Number of promoters in Learning Set | Number of promoters in Test Set 1 | Number of promoters in Test Set 2 |
|-----------|------------------|-------------------------------------|------------------------------------|------------------------------------|
|           |                  | TATA  | TATA  | Total | TATA  | TATA  | Total | TATA  | TATA  | Total |
| ppdb      | Arabidopsis      | 1,199 | 1,653 | 2,852 | 175   | 325   | 500   | 0     | 0     | 0     |
|           | Rice             | 1,899 | 6,630 | 8,529 | 103   | 397   | 500   | 0     | 0     | 0     |
| PlantProm DB | Dicots¹         | 178   | 90    | 268   | 0     | 0     | 0     | 12    | 12    | 24    |
|           | Arabidopsis      | 40    | 45    | 85    | 0     | 0     | 0     | 10    | 11    | 21    |
|           | Monocots²        | 48    | 42    | 90    | 0     | 0     | 0     | 4     | 2     | 6     |
|           | Rice             | 26    | 20    | 46    | 0     | 0     | 0     | 2     | 2     | 4     |
|           | Other³           | 5     | 18    | 23    | 0     | 0     | 0     | 0     | 0     | 0     |
|           | **Total**        | **3,395** | **8,498** | **11,893** | **178** | **722** | **1,000** | **28** | **27** | **55** |

¹ Except *A. thaliana*. ² Except *O. sativa*. ³ Other species: *Chlamydomonas reinhardtii, Chlorella vulgaris, Chlorococcum littorale, Dunaliella tertiolecta, Pinus sylvestris, Pinus thunbergii, Pseudotsuga menziesii, Pyrenomonas salina, Scenedesmus obliquus* and *Volvox carteri* (for details see: Table S2).
Table S2. List of species represented in PlantProm DB and classification of promoters

| Species                                      | Taxon | Number of TATA promoters | Number of TATA-less promoters | Total number of promoters |
|----------------------------------------------|-------|---------------------------|-------------------------------|---------------------------|
| Aegilops tauschii                           | Monocot | -                         | 1                             | 1                         |
| Avena fatua                                  | Monocot | 2                         | -                             | 2                         |
| Avena sativa                                 | Monocot | 2                         | -                             | 2                         |
| Dendrobium grex Madame Thong-IN              | Monocot | -                         | 3                             | 3                         |
| Hordeum vulgare                              | Monocot | 19                        | 13                            | 32                        |
| Lophopyrum elongatum                         | Monocot | -                         | 1                             | 1                         |
| Musa x paradisiaca                           | Monocot | -                         | 1                             | 1                         |
| Oryza sativa                                 | Monocot | 30                        | 22                            | 52                        |
| Sauromatum guttatum                          | Monocot | 1                         | -                             | 1                         |
| Triticum aestivium                           | Monocot | 14                        | 5                             | 19                        |
| Zea mays                                     | Monocot | 16                        | 20                            | 36                        |
| Actinidia deliciosa                          | Dicot  | 1                         | -                             | 1                         |
| Antirrhinum majus                            | Dicot  | 2                         | 1                             | 3                         |
| Arabidopsis thaliana                         | Dicot  | 52                        | 57                            | 109                       |
| Atropa belladonna                             | Dicot  | 1                         | -                             | 1                         |
| 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                         |
| Citrus sinensis                              | Dicot  | -                         | 1                             | 1                         |
| Craterostigma plantagineum                   | Dicot  | 2                         | 4                             | 6                         |
| Cucumis sativus                              | Dicot  | 3                         | -                             | 3                         |
| Daucus carota                                | Dicot  | 3                         | -                             | 3                         |
| Dianthus caryophyllus                        | Dicot  | 2                         | 1                             | 3                         |
| Eucalyptus gunnii                            | Dicot  | 2                         | -                             | 2                         |
| Flaveria brownie                             | Dicot  | 1                         | -                             | 1                         |
| Flaveria trinervia                           | Dicot  | 1                         | -                             | 1                         |
| Glycine max                                  | Dicot  | 21                        | 10                            | 31                        |
| Gossypium hirsutum                           | Dicot  | 6                         | -                             | 7                         |
| Helianthus annuus                            | Dicot  | 7                         | 1                             | 8                         |
| Hevea brasiliensis                           | Dicot  | 1                         | -                             | 1                         |
| Ipomoea batatas                              | Dicot  | 2                         | -                             | 2                         |
| Ipomoea nil                                  | Dicot  | -                         | 1                             | 2                         |
| Linum usitatissimum                          | Dicot  | 1                         | -                             | 1                         |
| Lotus japonicus                              | Dicot  | 2                         | -                             | 2                         |
| Lupinus luteus                                | Dicot  | 2                         | -                             | 2                         |
| Lycopersicon esculentum                      | Dicot  | 17                        | 8                             | 25                        |
| Lycopersicon penellii                        | Dicot  | 1                         | -                             | 1                         |
| Lycopersicon peruvianum                      | Dicot  | 1                         | -                             | 1                         |
| Lycopersicon pimpinellifolium                | Dicot  | 1                         | -                             | 1                         |
| Madagascar periwinkle                         | Dicot  | 1                         | -                             | 1                         |
Table S2 (continued)

| Plant Name                        | Taxonomic Class | 1 | 2 | 3 | 4 | 5 |
|-----------------------------------|-----------------|---|---|---|---|---|
| Malus domestica                   | Dicot           |   | 3 | - | 3 |   |
| Medicago sativa                  | Dicot           |   | 2 | 1 | 3 |   |
| Medicago truncatula              | Dicot           |   | 1 | - | 1 |   |
| Nicotiana benthamiana            | Dicot           |   | 2 | - | 2 |   |
| Nicotiana glutinosa              | Dicot           |   | 4 | 2 | 6 |   |
| Nicotiana plumbaginifolia        | Dicot           |   | 7 | 2 | 9 |   |
| Nicotiana tabacum                | Dicot           |   | 21| 10| 31|   |
| Papaver somniferum               | Dicot           |   | 3 | - | 3 |   |
| Petroselinum crispum             | Dicot           |   | 3 | 5 | 8 |   |
| Petunia hybrida                  | Dicot           |   | 6 | 2 | 8 |   |
| Petunia integrifolia             | Dicot           |   | 1 | - | 1 |   |
| Pharbitis nil                    | Dicot           |   | 1 | - | 1 |   |
| Phaseolus coccineus              | Dicot           |   | 1 | - | 1 |   |
| Phaseolus vulgaris               | Dicot           |   | 9 | 3 | 12|   |
| Pisum sativum                    | Dicot           |   | 20| 11| 31|   |
| Populus balsamifera subsp.       | Dicot           |   | - | 1 | 1 |   |
| Populus kitakamiensis            | Dicot           |   | - | 2 | 2 |   |
| Prunus persica                   | Dicot           |   | 1 | - | 1 |   |
| Psophocarpus tetragonolobus      | Dicot           |   | 1 | - | 1 |   |
| Pyrus serotina                   | Dicot           |   | 1 | - | 1 |   |
| Ricinus communis                 | Dicot           |   | 1 | 1 | 2 |   |
| Sesbania rostrata                | Dicot           |   | 1 | - | 1 |   |
| Sinapis alba                     | Dicot           |   | 1 | - | 1 |   |
| Solanum commersonii              | Dicot           |   | 1 | 1 | 2 |   |
| Solanum melongena                | Dicot           |   | 2 | - | 2 |   |
| Solanum tuberosum                | Dicot           |   | 9 | 5 | 14|   |
| Spinacia oleracea                | Dicot           |   | - | 7 | 7 |   |
| Striga hermonthica               | Dicot           |   | - | 1 | 1 |   |
| Vicia faba                       | Dicot           |   | 2 | 1 | 3 |   |
| Vigna radiate                    | Dicot           |   | 1 | - | 1 |   |
| Pinus sylvestris                 | Conifers        |   | 1 | 1 | 2 |   |
| Pinus thunbergii                 | Conifers        |   | 1 | - | 1 |   |
| Volvox carteri                   | Chlorophyta     |   | - | 2 | 2 |   |
| Pseudotsuga menziesii            | Conifers        |   | 1 | - | 1 |   |
| Pyrenomonas salina               | Cryptophyta     |   | - | 1 | 1 |   |
| Scenedesmus obliquus             | Chlorophyta     |   | - | 2 | 2 |   |
| Dunaliella tertiolecta           | Green algae     |   | - | 2 | 2 |   |
| Chlamydomonas reinhardtii        | Chlorophyta     |   | 2 | 8 | 10|   |
| Chlorrella vulgaris              | Chlorophyta     |   | - | 1 | 1 |   |
| Chlorococccum littorale          | Chlorophyta     |   | - | 1 | 1 |   |

1For the gene content of PlantProm DB visit: [http://www.softberry.com/data/plantprom/Links/Taxon_Table_2.htm](http://www.softberry.com/data/plantprom/Links/Taxon_Table_2.htm).
**Table S3.** Negative sets used in training and testing procedures

| Species    | Number of sequences in Learning Set | Number of sequences in Test Set 1 |
|------------|------------------------------------|-----------------------------------|
| TATA       | TATA        | TATA'               | TATA         | TATA'               |
| Arabidopsis| 3,533       | 8,858               | 278          | 861                 |
| Rice       | 2,467       | 6,142               | 222          | 639                 |
| **Total**  | **6,000**   | **15,000**          | **500**      | **1,500**           |

TATA' indicates TATA-less promoters.

---

**EXPECTATION MAXIMIZATION ALGORITHM**

In a simple implementation of Expectation Maximization (EM) algorithm (1), a sequence of a motif \(X=(x_1,x_2,\ldots,x_l)\) is considered, where \(l\) is a motif length. If \(p_i(x_j)\) is an empiric frequency of a nucleotide \(x_j\) (A, C, G, G) in position \(i\) (computed in previous iteration), then weight of such motif is computed as

\[W(X) = \frac{(\log \prod p_i(x_j))}{0.25}\]

---

**COMPUTATION OF FEATURE WEIGHTS (SCORES)**

To compute weight for features used to distinguish between promoter and non-promoter sequences, the following approach was used. Assume that a given promoter or non-promoter sequence is \(Y=(y_1,y_2,\ldots,y_L)\), where \(L\) is a sequence length, \(y_j\) is nucleotide (A or C or G or T) in position \(i\) (computed in previous iteration). Let’s also assume that \(q_1,q_2,q_3\) and \(q_4\) are frequencies of A, C, G and T in a sequence and a feature weight is computed for the search region \([T_1:T_2]\), where \(1 \leq T_1 \leq L\), \(1 \leq T_2 \leq L\).

**Computing weights for TATA-box, INR, YP and DPE motifs.** Consider sequence of a motif \(X=(x_1,x_2,\ldots,x_l)\), where \(l\) is a motif length and assume that \(p_i(x_j)\) is the empirical frequency of nucleotide \(x_j\) (A, C, G, G) in position \(i\), which is given by the corresponding Nucleotide Frequency Matrix (NFM). For a given \([T_1:T_2]\) search region of length \(T\) (\(T \leq l\)), \(K\) localizations \([t_1:t_2]\) for a motif are possible, where \(T=T_2-T_1+1\), \(k=T-l+1\). First, for every motif \(X_k\) \((k=1,2,\ldots,K)\) in \([t_1:t_2]\), weight is computed as

\[w_k(X_k) = \frac{(\log \prod p_i(x_j))}{q_j}\]

Then, the highest weight \(W(X)\) for motif \(X\) is found as

\[W(X)=\max(w_1,w_2,\ldots,w_K)\]

Search for putative TATA-box, INR and YP motifs in TATA promoter sequences was performed in \([-42:-19]\), \([-14:+14]\) and \([-50:+35]\) regions, respectively (+1 corresponds to the TSS); search for putative INR, YO and DPE motifs in TATA-less promoter sequences was performed in \([-14:+14]\), \([-50:+1]\) and \([+20:+35]\) regions, respectively.

**Computing scores for \(d(TATA,TSS)\) and \(d(TSS,INR)\) distances.** To obtain a frequency histogram (FH) for \(d(TATA,TSS)\), distance between TATA-box and TSS, and \(d(TATA,INR)\), distance between TATA-box and INR, for every entry from a positive learning set of TATA promoters we calculated distances between the following elements: \(d(TATA,TSS)\) and \(d(TATA,INR)\), which varied from 19 to 35 and from 0 to 14 bp,
respectively. Then, for every value of distance, (19-35) for d(TATA,TSS) and (0-14) for d(TATA,INR), we calculated frequency of its occurrence in sequences of a set and compiled the TATA-TSS and TATA-INR FHs. To get a score of d(TATA,TSS) for a given sequence, we calculated a distance between TATA-box and TSS, and for a score, we took the corresponding frequency from the TATA-TSS FH. To get a score of d(TATA,INR) for a given sequence, we calculated a distance between TATA-box and INR, and for a score, we took the corresponding frequency from the TATA-INR FH.

**Computing oligomer scores.** First, for positive and negative learning sets, we computed separate frequency tables for 2-mers and 3-mers in [-21:+20] region, for 4-mers, 5-mers and 6-mers in [-1:-200] and [+1:+51] regions. The final value in oligomer score table (OST) for a given k-mer was computed as a ratio of corresponding oligomer frequencies in positive and negative sets. To compute a score for k-mers (k=2,3,4,5) in a given sequence, to every k-mer, available in the corresponding search region, we assigned the corresponding score from the FH. Total k-mer score was calculated as a sum of these frequencies.

**Computing scores for TFBS density.** Data on transcription factor binding sites (TFBSs) were obtained from Regsite DB (www.softberry.com; Plant division) that contained 1,976 TFBSs. In both positive and negative learning sets, within [-200:+1] region of every sequence, we performed two kinds of search for motifs with similarity level 80% or higher: (1) search for motifs only in the sense strand of DNA; (2) search for motifs in both strands. We selected only TFBS motifs found in ≥5% of positive entries. We then, calculated frequencies of occurrence of each TFBS motif, separately in a positive and in a negative set, and as a final frequency for the corresponding TFBS, we took a ratio of corresponding frequencies in positive and negative sets. This way, we obtained two sets of TFBS and two corresponding TFBS score tables. To compute scores for TFBS density 1 and TFBS density 2 for a given sequence, within [-200:+1] region we performed search for TFBSs from two sets mentioned above. A total score for TFBS density 1 and 2 was computed as a sum of scores of TFBS found on a sense strand and on both strands, respectively.

**Computing scores for CG skew and AC skew.** To score the asymmetric nucleotide composition for a given DNA sequence, such as CG skew, \( sk(CG) \) and AC skew, \( sk(AC) \), we applied the following formulas, while number of nucleotides A (a), C (c), G (g) and T (t) in [-200:+20] region is given (2):

\[
sk(CG) = \frac{(c-g)}{(c+g)} \\
sk(AC) = \frac{(a+c-g-t)}{(a+c+g+t)}
\]

---

**NEURAL NETWORK TECHNIQUE AND VISAN SOFTWARE**

We used Neural Networks (NN) to get a recognition function for distinguishing between promoter and non-promoter sequences. The implementation used was part of VISAN software package (http://www.softberry.com/berry.phtml?topic=fdp.htm&no_menu=on).

VISAN is a software tool that provides a broad collection of machine learning algorithms. It allows users to quickly try out and compare different machine learning methods on new data sets. The machine learning algorithms include discriminant analysis, SVM, neural networks, random forests, and AdaBoost. The data and results can be visualized; both 2D and 3D visualization is possible. Statistical analysis of data is available as well. The main focus of VISAN is ease of use. It has a graphical user interface that enables easy access to the underlying functionality. Most tasks can be performed with a few clicks. The performance of different algorithms implemented in VISAN can be compared using accuracy histograms. We chose neural networks for our data analysis.
Neural networks are supervised learning algorithms inspired by biological neural networks, which can approximate any non-linear function. Supervised learning is a machine-learning task in which data labels are known. Neural networks consist of interconnected layers of neurons. The first layer is called an input layer. Its number of neurons is equal to the number of features in the data. The last layer is called an output layer and its number of neurons depends on the number of different data labels. The layers between input layer and output layer are called hidden layers. For the task at hand, we chose network architecture with one hidden layer of 15 neurons and two neurons for the output layer, as illustrated in Figure S1.

The neural network is trained using backpropagation algorithm, which calculates gradient of an error function with respect to weights in the network (for details of neural network based approaches see [2]). Each connection between neurons has an associated weight $W_{ij}$. Training a neural network involves tuning the weight corresponding to each connection. The weights are used to compute an output $a^l$ of each neuron that is called neuron activation:

$$a^l = f(W^{l-1}x + b^{l-1}),$$

where $l$ is the layer the neuron belongs to, $x$ is an input from the previous layer, $b^{l-1}$ is a bias, and $f$ is called an activation function that is sigmoid in our case.

The neural network is trained using backpropagation algorithm, which calculates gradient of an error function with respect to weights in the network [2]. After the model is trained, we can get a score for a new input object $o$ as a difference of values for neurons in the output layer:

$$F(o) = a^1_1 - a^1_2 + d,$$

where $a^1_1$ and $a^1_2$ are activations of output neuron 1 and 2 respectively for input $o$ and $d$ is threshold with default value 0.

STATISTICAL ESTIMATION OF PROMOTER PREDICTORS’ PERFORMANCE

To estimate performance of promoter predictors, we used the following statistical measures based on observed number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) predictions (5,6):

Recall (sensitivity or true positive rate, Sn): $Sn = TP/(TP+FN)$
Specificity (true negative rate, Sp): \( Sp = \frac{TN}{TN+FP} \)

F1-score (harmonic mean of Precision and Accuracy, F1): \( F1 = \frac{2TP}{2TP+FP+FN} \)

Mathew correlation coefficient (MCC):

\[
MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]
**Figure S2.** General scheme for computation of scores (weights) of promoter features used in a NN algorithm that was implemented into TSSPlant program.
**Figure S3.** Distribution of distances between the closest TSSpr and gene start (annotated TSS, TSSan) for 38,702 protein-coding genes of *G. max*.

**Figure S4.** Distribution of distances between the closest TSSpr and gene start (annotated TSS, TSSan) for 17,645 protein-coding genes of *P. trichocarpa*. 
**Figure S5.** Distribution of distances between the closest TSSpr and gene start (annotated TSS, TSSan) for 11,035 protein-coding genes of *V. vinifera*.

**Figure S6.** Relative location of TSSs mapped experimentally and predicted by TSSPlant program in upstream region of the AT2G41190 gene encoding a transmembrane amino acid transporter family protein. Chromosomal positions of TSSs correspond to TAIR *A. thaliana* genome annotation, ver6.0. The “unknown” indicates that promoter class for the corresponding TSS is unknown. Black arrows – experimental and grey arrows – predicted TSSs.
REFERENCES

1. Cardon, L.R. and Stormo, G.D. (1992) Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *J Mol Biol*, 223: 159-170.

2. Lobry, J.R. (1996) Asymmetric Substitution Patterns in the Two DNA Strands of Bacteria. *Mol Biol Evol.*, 13: 660-665.

3. Haykin, S. (1998) Neural Networks: A Comprehensive Foundation, Prentice Hall, 842 p.

4. Bengio, Y. (2009) Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning*, 2: 1-127.

5. Schneider, T.D., Stormo, G.D., Gold, L. and Ehrenfeucht, A. (1986) Information content of binding sites on nucleotide sequences. *J Mol Biol*, 188: 415-431.

6. Powers, D.M. (2011) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *J of Machine Learning Technologies*, 2: 37-63.