A peer-reviewed version of this preprint was published in PeerJ on 1 November 2019.

View the peer-reviewed version (peerj.com/articles/7990), which is the preferred citable publication unless you specifically need to cite this preprint.

As interest in genetic resequencing increases, so does the need for effective mathematical, computational, and statistical approaches. One of the difficult problems in genome annotation is determination of precise positions of transcription start sites. In this paper we present TransPrise - an efficient deep learning tool for prediction of positions of eukaryotic transcription start sites. TransPrise offers significant improvement over existing promoter-prediction methods. To illustrate this, we compared predictions of TransPrise with the TSSPlant approach for well annotated genome of *Oryza sativa*. Using a computer equipped with a graphics processing unit, the run time of TransPrise is 250 minutes on a genome of 374 Mb long. We provide the full basis for the comparison and encourage users to freely access a set of our computational tools to facilitate and streamline their own analyses. The ready-to-use Docker image with all necessary packages, models, code as well as the source code of the TransPrise algorithm are available at (http://compubioverne.group/). The source code is ready to use and customizable to predict TSS in any eukaryotic organism.
TransPrise: a novel machine learning approach for eukaryotic promoter prediction

Stepan Pachganov\textsuperscript{1}, Khalimat Murtazalieva\textsuperscript{2,3}, Alexei Zarubin\textsuperscript{4}, Dmitry Sokolov\textsuperscript{5}, Duane Chartier\textsuperscript{6}, Tatiana V. Tatarinova\textsuperscript{2,7-9}

\textsuperscript{1} Ugra Research Institute of Information Technologies, Khanty-Mansiysk, Russia
\textsuperscript{2} Vavilov Institute of General Genetics, Moscow, Russia
\textsuperscript{3} Institute of Bioinformatics, Moscow, Russia
\textsuperscript{4} Tomsk National Research Medical Center of the Russian Academy of Sciences, Research Institute of Medical Genetics, Tomsk, Russia
\textsuperscript{5} Neirika Solutions, Sarov, Russia
\textsuperscript{6} International Center for Art Intelligence, Inc. Los Angeles, CA, USA
\textsuperscript{7} Department of Biology, University of La Verne, La Verne, CA, USA
\textsuperscript{8} A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia
\textsuperscript{9} Siberian Federal University, Krasnoyarsk, Russia.

Abstract

As interest in genetic resequencing increases, so does the need for effective mathematical, computational, and statistical approaches. One of the difficult problems in genome annotation is determination of precise positions of transcription start sites. In this paper we present TransPrise - an efficient deep learning tool for prediction of positions of eukaryotic transcription start sites. TransPrise offers significant improvement over existing promoter-prediction methods. To illustrate this, we compared predictions of TransPrise with the TSSPlant approach for well annotated genome of \textit{Oryza sativa}. Using a computer equipped with a graphics processing unit, the run time of TransPrise is 250 minutes on a genome of 374 Mb long.

We provide the full basis for the comparison and encourage users to freely access a set of our computational tools to facilitate and streamline their own analyses. The ready-to-use Docker image with all necessary packages, models, code as well as the source code of the TransPrise algorithm are available at (http://compubioverne.group/). The source code is ready to use and customizable to predict TSS in any eukaryotic organism.
**Introduction**

The heightened interest and lowered costs of sequencing has led to intensified use of mathematical and statistical tools to improve the accuracy and replicability of genomic analysis. Until recently, there has been a relatively simplistic and deterministic notion of promoter organization and the creation of functional models of gene regulation. We are at a turning point in genetic research – we are actively assembling and refining the “dictionary” of genetics, but we are in the infancy of development of “genetic grammar”. Multiple questions remain to be addressed: what do the individual “words” mean? How does genotype translate to phenotype? This is very much like the time in physics, over 100 years ago, when the classical approach had to be replaced by quantum theory and statistical mechanics. We cannot understand genetics without powerful statistical and computational tools for analysis and for verifying our hypotheses regarding genome manifestations.

Thousands of eukaryotic genomes have been sequenced so far (https://www.ncbi.nlm.nih.gov/genome/browse/#!/eukaryotes/), including animals (1,590), fungi (3,275), and plants (665). As of 2018, these genomes are at various assembly levels, with 840 genomes assembled at the level of chromosomes, 46 are complete, 1,191 are in contigs, and 4,057 are at the level of genomic scaffolds. Genomic projects are not limited to sequencing and genome assembly. Re-sequencing large populations is becoming an important tool to unravel population structure, detect signatures of selection and to map quantitative trait loci (QTL) [1]. As resequencing costs plummet and technology platforms continue to expand throughput (e.g. Illumina NovoSeq), genomics communities are now contemplating the possibilities of resequencing entire germplasm collections to detect the vast majority of existing alleles and haplotypes. One essential requirement to capture allelic diversity is to have high-quality reference genomes that span the breadth of genomic diversity for mapping resequencing data.

Understanding the functional role of a given single-nucleotide polymorphism or a structural variant requires knowledge of its location with respect to coding and regulatory regions and the elements involved [2-5]. In addition, the regulatory role of a transcription factor binding site (TBFS) has been demonstrated to depend on the position of the TFBS with respect to the transcription start site (TSS) [6, 7]. Determination of the precise location of TSS is an essential preparatory step for motif discovery and reconstruction of gene regulatory networks [8]. The interaction of a vast number of proteins, multi-subunit complexes, and DNA binding sites make eukaryotic transcriptional regulation an extremely convoluted process [9]. Therefore, it is vitally important to have reliable methods for promoter prediction and analysis of regulatory elements if we are to enhance our capacity to engineer crops or to select therapeutic targets.
Homology-based prediction of coding regions is a relatively straightforward procedure [10]. Multiple tools and pipelines exist for finding positions and functions of genes, such as MAKER [11, 12], BREAKER [13], Augustus [14], GeneMarkHMM [15], FgeneSH [16], and many others. These pipelines achieve remarkably high accuracy in homology-based gene finding; however, homology between species does not necessarily extend beyond coding regions, and, therefore, accurate prediction of promoters is difficult. It has been reported that even state-of-the-art modern methods of promoter mapping are incapable of achieving 100% accuracy [5, 8, 17-25]. For example, current annotations of rice (MSU7) and maize (B73, 6a) contain 56K and 63K predicted genes, correspondingly [26], and for nearly two thirds of those genes, TSS is not identified precisely. Traditional deterministic approaches can predict only ~50% of promoters with one false positive promoter predicted every 700 - 1000 nt of the genome [27, 28]. This accuracy is insufficient to make reliable predictions, because we expect one promoter per 10,000-20,000 nt of a genome. More sophisticated tools, such as PromH [29] used conservation of promoter functional components between orthologous genes to improve prediction of TSS. PromH was able to predict TSS within 10 nt for 90% of the TATA+ promoters and for 40% of TATA- genes, but only if there are highly similar homologous sequences from closely related species. The TSSer algorithm [8] that combined positional frequency of 5’ EST/RNA-Seq matches on genomic DNA with gene models and did use unreliable homology arguments was able to accurately predict one transcription start site per locus. However, it is now accepted that alternative promoters are associated with differential expression in various tissues and chromatin states [30]. A nonparametric maximum likelihood approach, NPEST [24], can predict multiple TSSs per locus if 5’ EST/CAGE/mRNA data are available. Promoter sequences predicted by NPEST were demonstrated to be more accurate for the A. thaliana genome than sequences identified in several gold standard databases, such as TAIR, Plant Prom DB and Plant Promoter Database. However, it is difficult to identify TSS from RNA-Seq alone, since only 26% of genes display a maximum of the RNA-Seq coverage in the range [TSS-50, TSS +250], and only 60% of genes display this maximum in the range [TSS-50, TSS+550] [31]. Enough RNA-Seq and CAGE data is not available for all genomes of interest. Therefore, it is imperative to develop alternative strategies.

There are several factors complicating the process of TSS prediction, such as existence of multiple TSS per locus. Studies on mammalian and plant genomes have revealed that many eukaryotic genes are associated with multiple distinct promoters [22, 23, 32, 33]. Moreover, eukaryotic promoters are characterized by multiple TSSs and can be classified based on the distribution and utilization of their collective TSSs. Consequently, the association with several distinct promoters allows for a single gene to encode various protein isoforms [34].

In addition, performance of standard promoter identification in grasses and warm-blooded vertebrates is complicated by the existence of two classes of genes in those organisms: GC3 –rich and –poor ones (where GC3 is the fraction of Cs and Gs in the third position of codons). Nucleotide composition of GC3 –rich genes differs from GC3 –poor ones; they also have higher variability of
gene expression levels (resulting in fewer full-length mRNA support) [35-37]. Since a majority of the stress-related and tissue-specific genes are GC$_3$-rich [38], refinement of promoter prediction pipeline is an essential task.

Many genomic features are associated with the location of promoter: positional frequency of 5’ ESTs and RNA-Seq matches on genomic DNA, nucleotide distribution, DNA methylation, distribution of SNPs, characteristic regulatory elements. Incorporation of those data types allows accurate prediction of TSS. A recently developed tool, TSSPlant [39], based on the Expectation Maximization (EM) algorithm, achieves significantly higher accuracy compared to state-of-the art promoter prediction programs for both TATA-containing and TATA-less promoters. Triska et al. [4] presented a deep learning approach to characterize regions as promoters and non-promoters, achieving 99% accuracy in classification of 250 nt long regions. However, the question of the specific location of the TSS within these 250 nt long windows remains open.

This paper presents a novel, accurate, and data-type independent procedure for TSS prediction that can incorporate multiple data types. Our method is based on machine learning approach that is capable of uncovering intricate properties of promoter regions and achieving much higher accuracy than deterministic methods [4, 40]. Our novel method aims to identify the position of the start of transcription with the highest possible precision using nucleotide composition alone. The method can predict multiple transcription start sites per locus. It is data-type agnostic and can be extended to incorporate additional biological features. We present a set of computational tools, a user-friendly public interface and a curated online database to facilitate these analyses.

**Materials and Methods**

**Selection of genome annotation version**

We selected rice chromosomes and Genome Annotation release 7 (MSUv7, [http://rice.plantbiology.msu.edu](http://rice.plantbiology.msu.edu)). There are two commonly used annotations of rice: MSU [5, 41] and Fgenesh [42]. The Fgenesh gene prediction set contains 18,389 high quality (5’ full, with mRNA support) gene models, while the MSU gene prediction set contains 20,367 high quality gene models [5]. We used Fgenesh mRNA-based gene prediction models, since Fgenesh-annotated promoters have a more pronounced nucleotide consensus as compared to the promoters annotated by MSU [4]. Fgenesh was successfully used to annotate several plant genomes [38, 43-49]. Therefore, we selected the Fgenesh annotation as the gold standard for our analysis. To obtain the highest quality dataset, pseudogenes, transposable elements, and genes with 5’ UTR shorter than 20 nt or longer than 1000 nt have been excluded.

**Training, validation and test sets**
The procedure consists of two steps: classification (dividing the genome into “promoters” and “non-promoters”) and regression (finding the position of TSS inside the sequence identified as “promoter”). The genomic sequence of *Oryza sativa* had been divided into the testing and training sets. The training set contains three files:

1. **Training “non-promoter” dataset** contains sequences extracted from random genomic positions separated from experimentally validated transcription start sites by 2000 nt. This dataset contains mostly intergenic regions. All sequences are 2000 nt long.

2. **Training “promoter” dataset** contains sequences [TSS-1000; TSS+999] from the all chromosomes with length 2000 nt.

3. File with **indicators of TSS positions**, containing (2000×1) matrices that correspond to positions of biologically validated TSS in every training sequence (“1” TSS, “0” not TSS position).

The same set of files was created for the testing dataset. Since the procedure has multiple steps (classification and regression), training and testing sets were selected at each step of the method.

The following procedure was used to assemble the dataset for the **classification model**:

1) “Non-promoters”: ¼ of the examples chosen from the training “non-promoter” dataset, randomly selecting 512 long sequences from 2000 nt long regions.

2) “Promoters sans TSS”: ¼ of the examples were randomly selected from the training “promoter” dataset, making sure that the chosen 512 nt long fragment did not overlap the region [TSS-50, TSS+50].

3) “TSS vicinity”: ½ of the examples extracted from the training “promoter” dataset, containing only one TSS in a random position within the 512 nt long sequence, with a restriction that it should be in the [250, 450] fragment.

The dataset for the **regression model** was assembled using sequences that contain one validated TSSs in a randomly selected position of the [250, 450] fragment.

Datasets are the (512×4) nucleotide matrices M with 512 columns and 4 rows. The 1st row contains function $\delta(i,A)$ - it is equal to 1 if there is nucleotide “A” in the i-th position of the sequence and 0 otherwise. Similarly, 2nd, 3rd, and 4th rows correspond to nucleotides C, G and T.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>508</th>
<th>509</th>
<th>510</th>
<th>511</th>
<th>512</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>...</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Model Training
We implemented the Convolutional Neural Networks (CNN) using the Keras library for training (https://keras.io/).

Classification and Regression models training
The dataset for the classification contains equal numbers of positive and negative examples. The matrices (512×4) described above are input into the model. The CNN architecture (Fig. 1) started with four parallel convolutional layers (composed of 128 filters with 2, 4, 8 and 16 kernel sizes) ReLU, was used as activation function followed by concatenation. After concatenation layer we used convolution, batch normalization, max pooling layers twice. The First convolution had 128 filters and the second had 16. There were one kernel size and ReLU activation in both situations. To help regularize the model, we used the 0.5 Dropout technique. The signal is fed to two fully connected layers with ReLU activation functions consisted of 256 and 128 neurons, followed by batch normalization. The output layer had a sigmoid activation function.

Whole Genome Sequencing (WGS) processing
TransPrise also works with WGS data in fasta format. We prepared a Python script that with 250 minutes run time on a genome of 374 Mb long. It uses a sliding window for extraction of 512 nt long sequences with the step size 4 nt. If classification step identified a fragment as TSS-containing, this fragment will be passed to the regression step. Prediction vector is compared to the TSS model using L1 norm. The default value for the similarity threshold is 600, but it can be modified by users. The only restriction of our algorithm is that alternative TSSs must be more than 100 nt of each other.
For the purpose of the K-fold cross-validation, the dataset is randomly divided into K equal-size subsets. Of the K subsets, a single subset is retained as the validation data for testing the model, and the remaining (K-1) subsets are used for training. The cross-validation process is then repeated K times (the “folds”), with each of the K subsamples used exactly once as the validation data. Then, the results from K folds are averaged. The advantage of K-fold cross-validation is that all observations are used for both training and validation, and each observation is used for validation exactly once. We conducted 10-fold cross-validation (dataset was divided into training and validation sets in 9:1 ratio; validation set was used to avoid overfitting and find the optimal number of learning epochs). The ROC curves obtained in 10-fold cross-validation are presented in the “Results” section. We determined that the optimal number of learning epochs is five. After the model training, we tested our model using the test set and calculated Accuracy (Ac), Sensitivity (Se), Specificity (Sp), and the Matthews Correlation Coefficient (CC):

\[
\text{Specificity} = \frac{TP}{TP + FP},
\]

\[
\text{Sensitivity} = \frac{TP}{TP + FN},
\]

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},
\]

\[
CC = \frac{TP \times TN + FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}},
\]

where TP - true positive, TN - true negative, FP - false positive, FN - false negative. The input of the regression model has the same format as the classification model input. TSS is assumed to be located at a random position between nucleotides 250 and 450. There is only one difference between classification and regression models: in the output layer the activation function is replaced by a linear function.

We trained the classification and regression models using the Keras Library and performed 10-fold cross validation procedures for them. The algorithm for TSS prediction is presented in the Fig.1. For every fold, we carried out five learning epochs. the complete learning time was 35 seconds on average. We performed the 10-fold cross-validation and calculated the average value of mean absolute error (MAE) to estimate the accuracy of TSS position prediction, where \( y_i \) - position of TSS in test set (assumed to be accurate), and \( x_i \) - predicted position of TSS.

\[
\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - x_i|,
\]

Fig. 2 shows the flowchart for model building and validation.
Results

Classification

To reduce the influence of how the data is split on the resulting testing statistics, we carried out a 10-fold cross-validation (CV) procedure for internal validation and supplemented it by external validation on the chromosome 2, which had been excluded from training set. We randomly divided the dataset into 10 partitions and used each partition as the testing data while training the model on the remaining partitions. Then we selected the best model and evaluated performance on chromosome 2 (external validation dataset). We performed external validation on the test set (containing of 2000 nucleotide sequences with length 512 nt, where 1000 examples - “non-TSS” sequences and 1000 - “TSS” sequences) for TransPrise and TSSPlant [4, 39, 40] classification models and calculated the Matthews correlation coefficient (MCC), Accuracy (Ac), Sensitivity (Se), Specificity (Sp), and Area Under the ROC Curve (AUC-ROC). The ROC curve (Receiver Operating Characteristic curve) represents the dependence of a sensitivity on the specificity, or alternatively, is a graph showing the performance of a classification model at all classification thresholds. The results are presented in the Table 1.

Table 1: Comparison of accuracy metrics of TransPrise and TSSPlant classification models

AUC-ROC curves for classifier obtained in 10-fold cross-validation are presented in the Figure 3. Accuracy = 0.88, Se = 0.84, Sp = 0.92, CC = 0.79, AUC = 0.94.
Regression

Figure 4 shows the error density curves obtained in the 10-fold cross validation procedure for regression models. The mean absolute error (MAE) for regression model was 47.68 nt.

We compared the accuracy of promoter prediction for TSSPlant and TransPrise. For this purpose, we selected rice chromosomes 1 and 2. The testing was performed for entire chromosomes, without filtering for possible TSS containing regions, located upstream of the ATG.

These two chromosomes contain 5,298 high quality experimentally validated TSS, defined as follows:

1) Locus does not correspond to transposable element
2) Locus has experimental support (full-length mRNA)
3) If multiple isoforms are predicted, the “representative” is used
4) Size of the 5’ UTR is at least 20 nt.

For these two chromosomes, TSSPlant predicted 153,009 TSSs, while TransPrise has found 13,765 sites. Of TSSPlant predictions, 10,721 (~7%) were located within 1000 nt from validated TSS. Of TransPrise predictions, 3,989 (~29%) were located within 1000 nt from validated TSS. Additionally, TransPrise predictions tend to be closer to validated TSS than TSSPlant predictions (Figure 5).

Discussion

We have developed an efficient, deep learning approach, for prediction of the position of transcription start sites in eukaryotes using properties of a nucleotide sequence. The approach is data-type independent and allows incorporation of additional data types (such as RNA-seq and tissue specific DNA methylation), refining positions of TSS for tissue-specific and stress-specific expression.

We compared TransPrise with the TSSPlant approach on an independent test set composed of 2000 nucleotide sequences. All sequences were 512 nt long, and 1000 sequences did not contain TSS (“non-TSS”), and 1000 contained TSS (“TSS” sequences). The Matthews correlation coefficient value for TransPrise is more than twice larger than for TSSPlant classification models (0.79 vs. 0.31), indicating the significantly higher efficiency of TransPrise in distinguishing between regions that contain and do not contain starts of transcription. Additionally, a regression model was created for precise localization of TSS within the sequence classified as a “promoter”. We validated our regression model on a test set composed of 1000 “TSS” sequences selected from chromosome 2 and calculated the mean absolute error to be 47 nt.
Another important genome annotation task is identification of functional motifs. The architecture of TransPrise is especially designed for that. The first convolution layer is composed of four different kernel size filters \((a_{ij}) - 4*2, 4*4, 4*8, 4*16\) matrices, where \([i:1,0,0,0](A),[0,1,0,0](T),[0,0,1,0](C),[0,0,0,1](G)\) and \([j:2,4,8,16 - length of motif sequence]\) (in total 128 filters of each type). After model training, the filters correspond to PWM (position-specific weight matrix) describing informative sequences in promoters and can be visualized as sequence logos. Several filter motifs correspond to known regulatory elements: TGGGCC [50], CGATT [51], ACTCAT [52], and CGCG box. Motif TGGGCC is targeted by the TCP transcription factor through its interaction with proliferating cell nuclear antigens PCF1 and PCF2 [50]; ACTCAT motif [53] is a typical binding site of basic leucine zipper (bZIP) transcription factor; CGCG cis-elements are found in promoters of stress-related genes, for example involved in ethylene signaling, abscisic acid signaling, and light signal perception. They are bound by AtSR1 transcription factor [54].

Figure 6 shows “filter” motifs that correspond to two well-characterized features of eukaryotic promoters: Initiator element CA and TATA-box [4, 8, 55, 56]. Therefore, we have shown that at least some of the features selected by the model as informative for identification of TSS correspond to known, biologically validated regulatory elements, over-represented at or near the start of transcription. We believe that other features may correspond to unknown regulatory elements.

TransPrise is an advanced tool for improving genome annotation. This is symbolic of a mathematical approach to biology, which is increasingly significant. As the pressure to annotate genomes increases with plummeting sequencing costs, we must focus efforts on understanding the significance of individual genomic regions. TransPrise can predict important regulatory regions and identify characteristic motifs at a high level of efficiency relative to existing approaches. TransPrise is data-type and species independent tool that can be easily installed and customized.

**Software availability**

We offer a simple and efficient way to deploy our training models for most users’ devices without having to install third-party deep learning packages. We have implemented the ready-to-use Docker image with all necessary packages, models and code (http://compubioverne.group/). The source code for our program TransPrise is also available on GitHub (https://github.com/StepanAbstro/TransPrise) and is ready to use for new model training.


27. Shahmuradov, I.A. and V.V. Solovyev, Nsite, NsiteH and NsiteM computer tools for studying transcription regulatory elements: Fig. 1. Bioinformatics, 2015. 31(21): p. 3544-3545.


30. Rye, M., et al., Chromatin states reveal functional associations for globally defined transcription start sites in four human cell lines. BMC Genomics, 2014. 15: p. 120.


32. Louzada, E.S., Alternative Processing as a Mechanism for Regulating Gene Expression, in Regulation of Gene Expression in Plants. p. 67-100.


Figure 1 (on next page)

TransPrise algorithm. Inset: CNN architecture that was used in classification/regression model training

CNN architecture was used in classification and regression steps
Figure 2

Flowchart of model building and validation.

The figure presents a flowchart of the model building and validation.
Figure 3 (on next page)

ROC curve for the training dataset

ROC-curves obtained 10-fold cross-validation procedure of classification model. A: training dataset, B: testing dataset. Accuracy = 0.88, Se = 0.84, Sp = 0.92, CC = 0.79, AUC = 0.94
Figure 4 (on next page)

ROC curve for the testing dataset

ROC-curves obtained 10-fold cross-validation procedure of classification model. A: training dataset, B: testing dataset. Accuracy = 0.88, Se = 0.84, Sp = 0.92, CC = 0.79, AUC = 0.94
Figure 5 (on next page)

Error density curves obtained in 10-fold cross-validation of regression models.

The error density curves were obtained in the 10-fold cross validation procedure for regression models. The mean absolute error (MAE) for regression model was 47.68 nt. The testing was performed for entire chromosomes, without filtering for possible TSS containing regions, located upstream of the ATG.
Figure 6

Comparison between TransPrise and TSSPlant predictions

TransPrise predictions tend to be closer to validated TSS than TSSPlant predictions
Figure 7

Comparison between TransPrise and TSSPlant predictions (fraction of sites)

TransPrise predictions tend to be closer to validated TSS than TSSPlant predictions
Figure 8

Distribution of CA

CA motif is associated with the start of transcription, its density peaks at TSS
Figure 9

Distribution of TATA

TATA motif is frequently located at position -30, its density shows a peak at -30.
Table 1: Comparison of accuracy metrics of TransPrise and TSSPlant classification models

TransPrise offers higher accuracy compared to the TSSPlant
Table 1: Comparison of accuracy metrics of TransPrise and TSSPlant classification models

<table>
<thead>
<tr>
<th>Classification Model</th>
<th>MCC</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSSPlant</td>
<td>0.310</td>
<td>0.603</td>
<td>0.976</td>
<td>0.231</td>
<td>0.603</td>
</tr>
<tr>
<td>TransPrise</td>
<td>0.791</td>
<td>0.895</td>
<td>0.872</td>
<td>0.919</td>
<td>0.952</td>
</tr>
</tbody>
</table>