TransPrise – a machine learning approach for prediction of eukaryotic transcription start sites

Stepan Pachganov[1], Khalimat Murtazalieva[2], Alexei Zarubin[3], Duane Chartier[4], Tatiana V. Tatarinova[2,5-7]

1 Ugra Research Institute of Information Technologies, Khanty-Mansiysk, Russia
2 Vavilov Institute of General Genetics, Moscow, Russia
3 Tomsk National Research Medical Center of the Russian Academy of Sciences, Research Institute of Medical Genetics, Tomsk
4 ICAI, Inc. Los Angeles, CA, USA
5 Department of Biology, University of La Verne, La Verne, CA, USA
6 A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia
7 Siberian Federal University, Krasnoyarsk, Russia.

Abstract

TransPrise is an effective and efficient deep learning tool that significantly improves prediction of eukaryotic transcription start sites. The performance of TransPrise was compared with the CNNProm approach using the well annotated genome of *Oryza sativa*. TransPrise predictions offer a significant improvement over other promoter-prediction methods. The run time of TransPrise is XXX minutes on a genome of XXX long.

We present 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 and code is available at https://hub.docker.com/r/zarubinaa/tss-rice/ The source code of the TransPrise algorithm, is available on GitHub (https://github.com/StepanAbstro/TransPrise) and is ready to use to be customized to predict TSS in any eukaryotic organism.

Introduction

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) (Atwell et al. 2010). 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 (Li et al. 2015; Mulder 2018)[,(Tatarinova et al. 2016; Triska et al. 2017)]. 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) (Berendzen et al. 2006; Pritsker et al. 2004). Determination of the precise location of TSS is an essential preparatory step for motif discovery and reconstruction of gene regulatory networks (Troukhan et al. 2009). The interaction of a vast number of proteins, multisubunit complexes, and DNA binding sites make eukaryotic transcriptional regulation an extremely convoluted process (Eckardt 2014). 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 (Keilwagen et al. 2018). Multiple tools and pipelines exist for finding positions and functions of genes, such as MAKER (Holt & Yandell 2011; Campbell et al. 2014), BREAKER (Hoff et al. 2015), Augustus (Stanke & Morgenstern 2005), GeneMarkHMM

(Stanke & Morgenstern 2005; Lukashin 1998), FgeneSH (Salamov & Solovyev 2000) , 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 (Alexandrov et al. 2009; Alexandrov et al. 2006) (Troukhan et al. 2009) (Alexandrov et al. 2009) (Carninci et al. 2006; Kawaji et al. 2006) (Kawaji et al. 2014; Morton et al. 2014; Batut et al. 2013) (Tatarinova, Kryshchenko, et al. 2013; Herbig et al. 2013) (Tatarinova, Kryshchenko, et al. 2013). For example, current annotations of rice (MSU7) and maize (B73, 6a) contain 56K and 63K predicted genes, correspondingly (Liseron-Monfils et al. 2013), and for nearly two thirds of those genes, TSS is not identified precisely (Liseron-Monfils et al. 2013; Tatarinova et al. 2016). Traditional deterministic approaches can predict only ~50% of promoters with one false positive promoter predicted every 700 - 1000 nt of the genome (Solovyev et al. 2010)[2](Shahmuradov & Solovyev 2015). This accuracy is insufficient to make reliable predictions, because we expect one promoter occurrence per 10,000-20,000 nt of a genome. More sophisticated tools, such as PromH (Solovyev & Shahmuradov 2003; Solovyev 2003) 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 (Troukhan et al. 2009) that combined positional frequency of 5' EST/RNA-Seq matches on genomic DNA with gene models and did not rely on 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 (Rye et al. 2014). A nonparametric maximum likelihood approach, NPEST (Tatarinova, Kryshchenko, et al. 2013), allowed for prediction of 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] (Steijger et al. 2013). Sufficient RNA-Seq and CAGE data is not available for all genomes of interest, and it therefore imperative to develop alternative strategies.

There are several factors complicating the process of TSS prediction. The first factor is existence of multiple TSS per locus. Studies on mammalian and plant genomes have revealed that many eukaryotic genes are associated with multiple distinct promoters (Batut et al. 2013; Morton et al. 2014; Louzada n.d.; Farrell & Bassett n.d.). 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 (Sandelin et al. 2007).

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: GC_3 –rich and –poor ones (where GC_3 is the fraction of Cs and Gs in the third position of codons). Nucleotide composition of GC_3 –rich genes differs from GC_3 –poor ones; they also have higher variability of gene expression levels (resulting in fewer full-length mRNA support) (Tatarinova, Elhaik, et al. 2013; Elhaik et al. 2014),(Elhaik & Tatarinova 2012). Since a majority of the stress-related and tissue-specific genes are GC_3-rich (Chan et al. 2017), 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 motifs, etc of tissue/stress specificity, and analysis of roles of each alternative promoter. Incorporation of those data types allows accurate prediction of TSS. A recently developed tool, TSSPlant (Shahmuradov et al. 2017), 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. (Triska et al. 2017) 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, data-type independent, procedure for TSS prediction that can incorporate multiple data types. Our method is based on machine learning, that is capable of uncovering intricate properties of promoter regions and achieving much higher accuracy than deterministic methods (Umarov & Solovyev 2017)[Triska et al. 2017]. Our novel method aims to identify the position of the start of transcription with the highest possible precision using nucleotide composition alone. 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 database to enable 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). There are two commonly used annotations of rice: MSU (Tatarinova et al. 2016; Kawahara et al. 2013) and Fgenesh (Zhang et al. 2008). 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 (Tatarinova et al. 2016). 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 (Triska et al. 2017). Fgenesh was successfully used to annotate a number of plant genomes (Chan et al. 2017; Ito et al. 2005; Yao et al. 2005; Davis et al. 2010; Sanusi et al. 2018; Sheshadri et al. 2018; Nasiri et al. 2013; Jiang et al. 2015). 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"). Out of twelve rice chromosomes, chromosome 2 was used for external validation and other chromosomes were used for training. 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 indicator function delta(i,A) - it is equal to 1 if there is nucleotide A in the ith position of the sequence and 0 otherwise. 2nd, 3rd, and 4th rows correspond to nucleotides C,G and T.

	1	***2***	***2***	***3***	***5***				***508***	***509***	***510***	***511***	***512***
A	1	1	0	0	0	...	...	...	0	0	0	0	0
C	0	0	0	0	0				0	1	0	1	1
G	0	0	0	1	0				0	0	1	0	0
T	0	0	1	0	1				1	0	0	0	0

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** model 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. First convolution had 128 filters and second had 16. There were 1 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 standard, 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.

We conducted 10-fold cross-validation (dataset was divided into training set and validation set in 9:1 ratio; validation set was used to avoid overfitting and find optimal number of learning epoch). The ROC curves obtained in 10-fold cross-validation are presented in the "Results" section. We determined that 5 learning epoches is optimal. After the model training, we tested our model using the test chromosome (chromosome 2) and calculated Accuracy, Sensitivity (Se), Specificity (Sp), and Matthews Correlation Coefficient (CC):

$$Sp = \frac{TP}{TP+FP},$$

$$Se = \frac{TP}{TP+FN},$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN},$$

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

where TP - true positive, TN - true negative, FP - false positive, FN - false negative.

The input of the regression model have the same shape. TSS located at a random position in between nucleotides 250 to 450. The regression model had only one difference in output layer, where activation function was replaced by linear.

We performed 10-fold cross-validation and calculated average value of mean absolute error (MAE) to estimate the accuracy of TSS position prediction (yi - position of TSS in test set, xi - predicted position of TSS). For every fold, we carried out five learning epochs, the complete learning time, on average, takes 35 seconds

$$MAE = \frac{\sum_{i=1}^{n} |yi - xi|}{n}$$

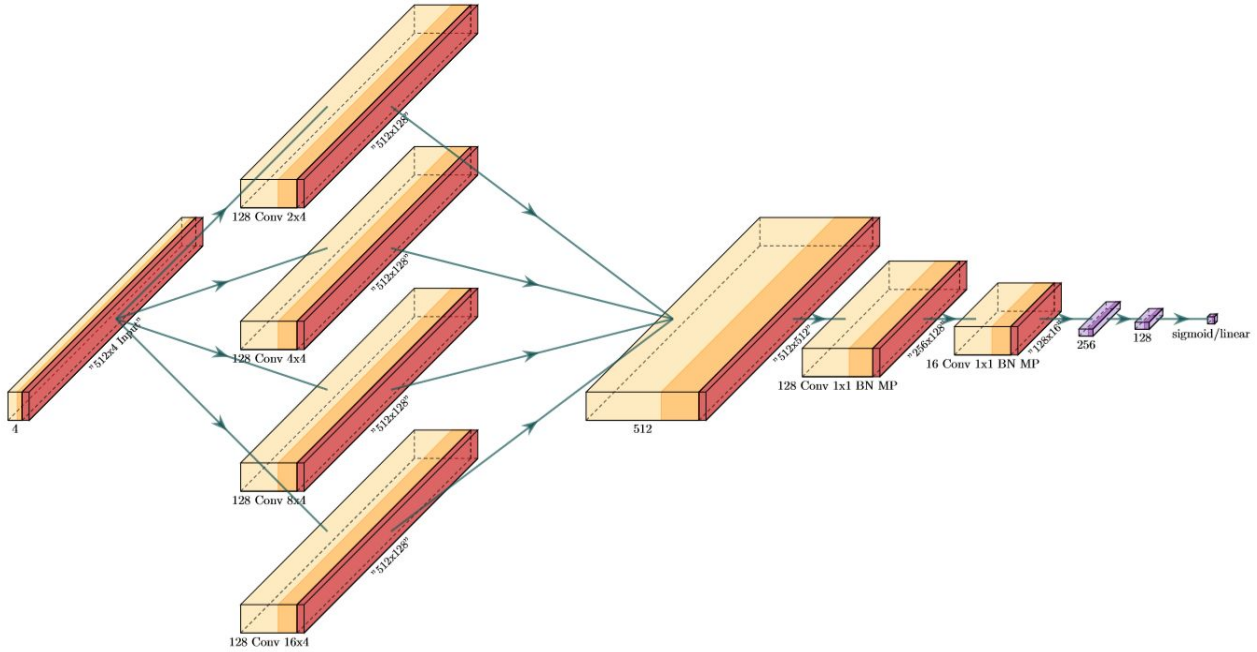
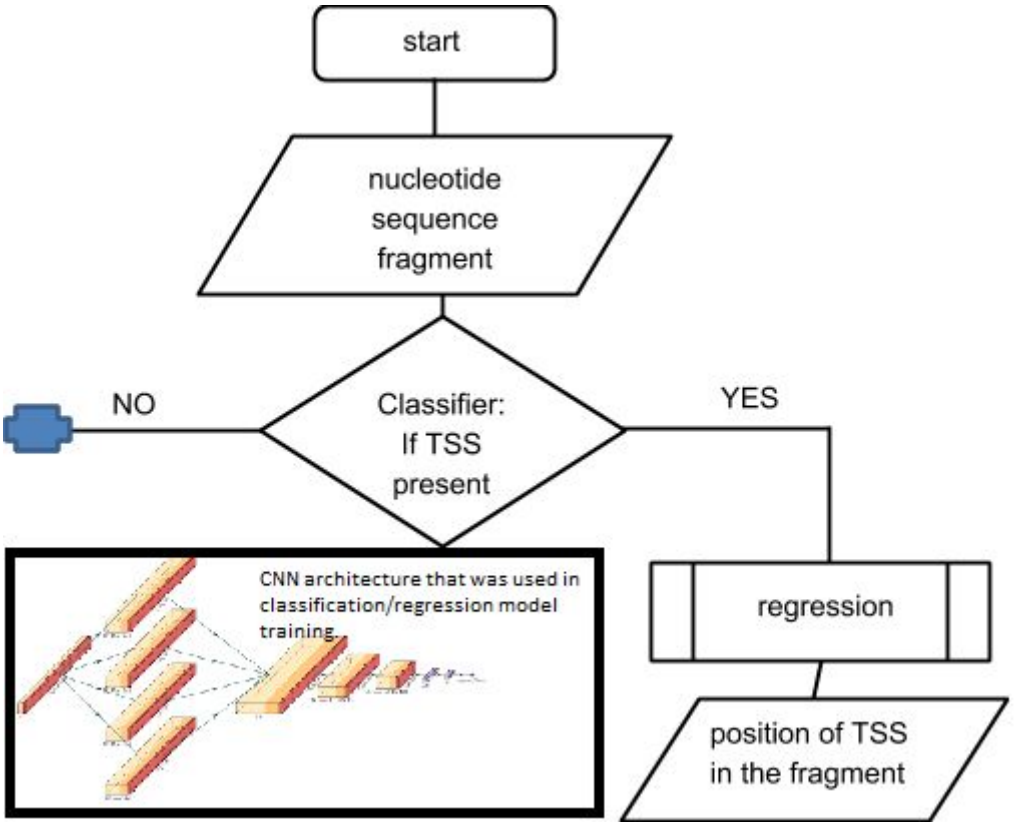


Fig 1. CNN architecture that was used in classification/regression model training.



The dataset had been divided into test (chromosome 2) and training set. The training set contains nucleotide sequences of genes. We trained classification and regression models using the Keras Library and performed 10-fold cross validation procedures for them. The algorithm for TSS predicting is presented in Fig.2.

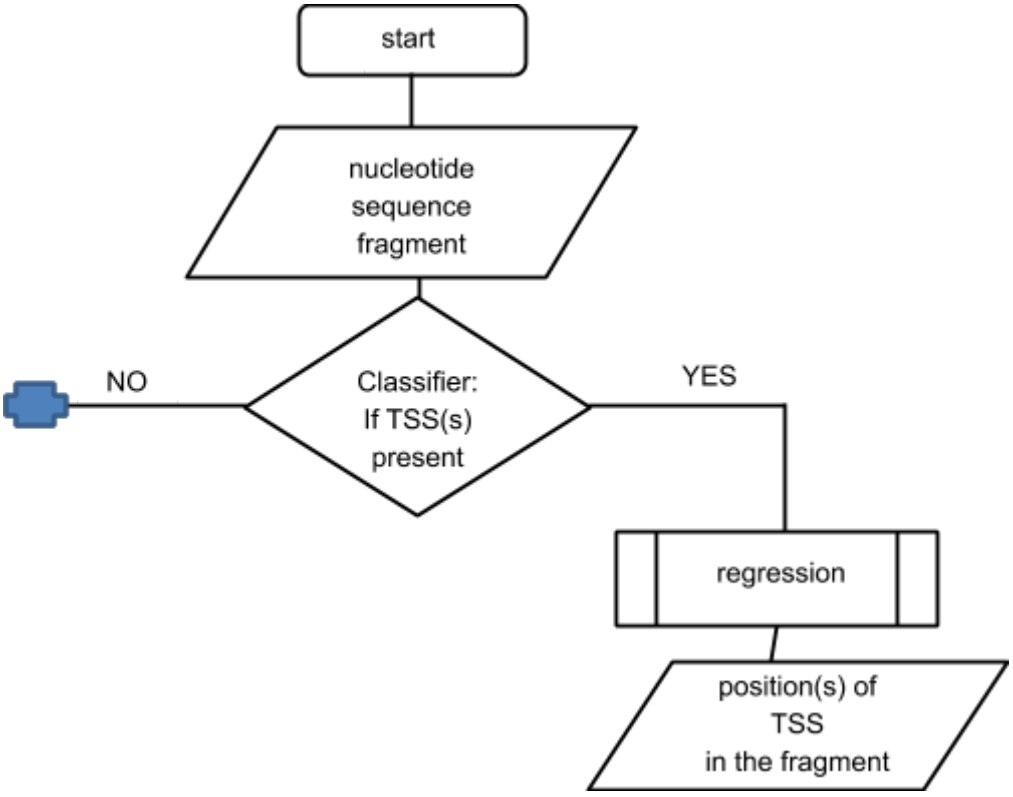


Fig. 2. The algorithm for TSS predicting.

Classification model

We performed external validation on the test set (composed of 2000 nucleotide sequences with length 512 nt from chromosome 2, where 1000 examples - "non-TSS" sequences and 1000 - "TSS" sequences) for TransPrize and CNNProm (Solovyev et al, 2017) classification models and calculated Matthews correlation coefficient (MCC), Accuracy, Sensitivity (Se), Specificity (Sp) and AUC (Area Under the ROC Curve). The result are presented in Table.1.

Table 1. Comparison of accuracy metrics of TransPrize and CNNProm classification models

Classification Model	**MCC**	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**

CNNProm	0.310	0.603	0.976	0.231	0.603
TransPrize	0.791	0.895	0.872	0.919	0.952

The ROC curve (Receiver Operating Characteristic curve) represents the dependance of a sensitivity on the specificity, or alternatively, is a graph showing the performance of a classification model at all classification thresholds.

AUC-ROC curves for classifier obtained in 10-fold cross-validation is presented in Fig.3. In 10-fold cross-validation, we randomly divided the original dataset into 10 equal-size subsamples. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. The cross-validation process is then repeated 10 times (the folds), with each of the 10 subsamples used exactly once as the validation data. Then, we averaged 10 results from the folds. The advantage of 10-fold cross-validation is that all observations are used for both training and validation, and each observation is used for validation exactly once.

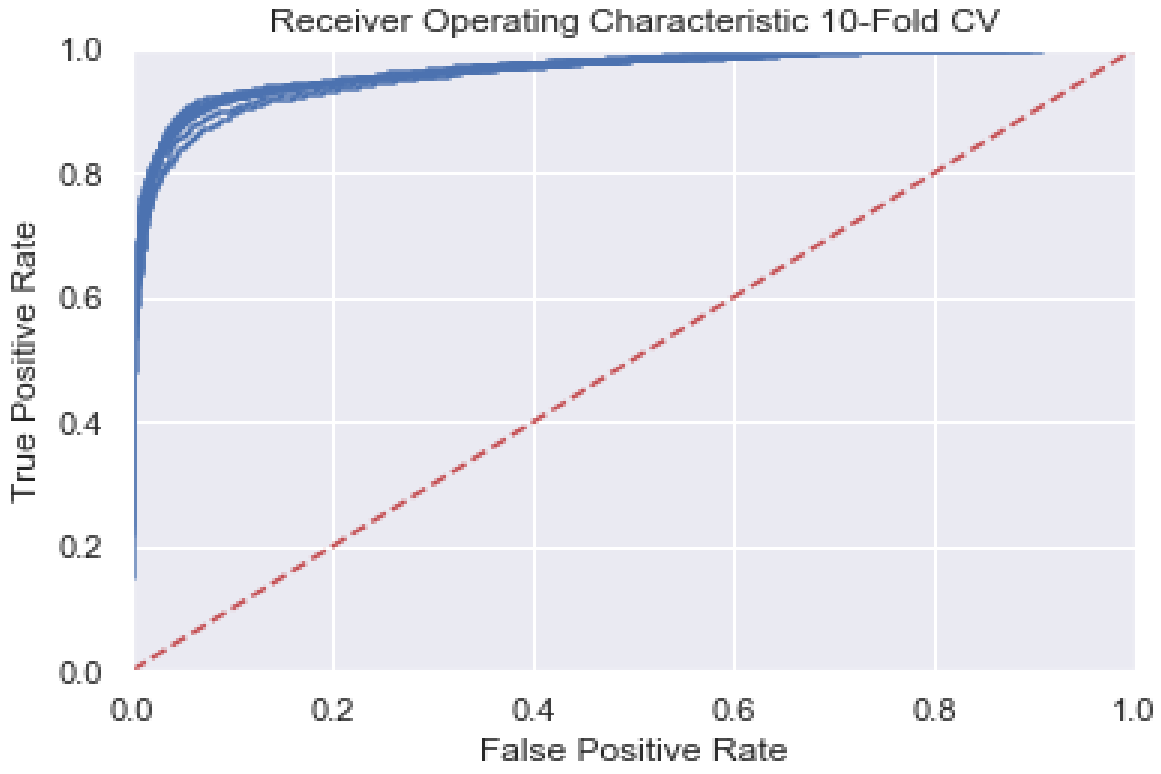


Fig. 3. ROC-curves obtained 10-fold cross-validation procedure of classification model.

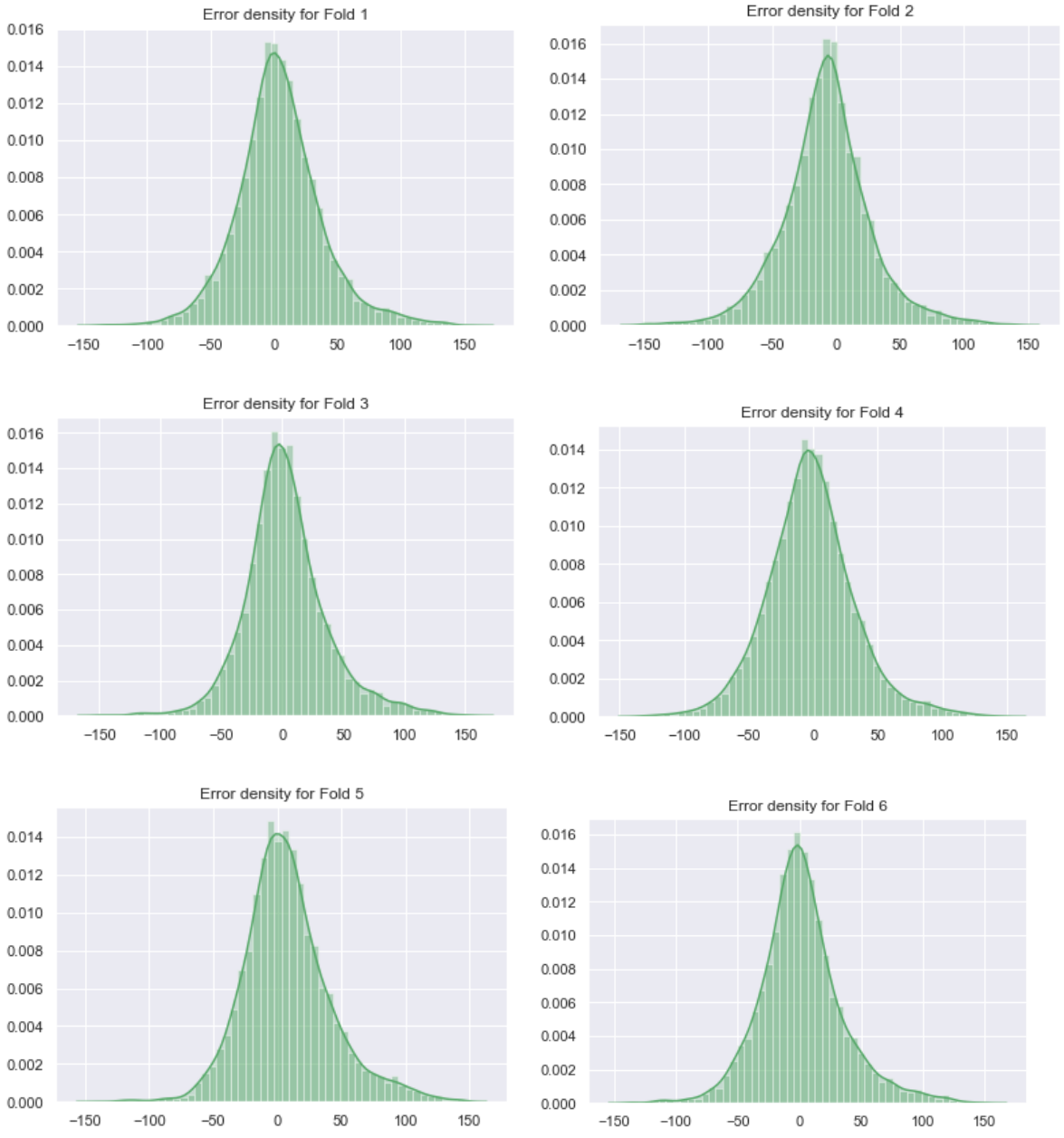
Also, we performed external validation of classification models on the test chromosome. The ROC-curves are presented in the Fig.4.

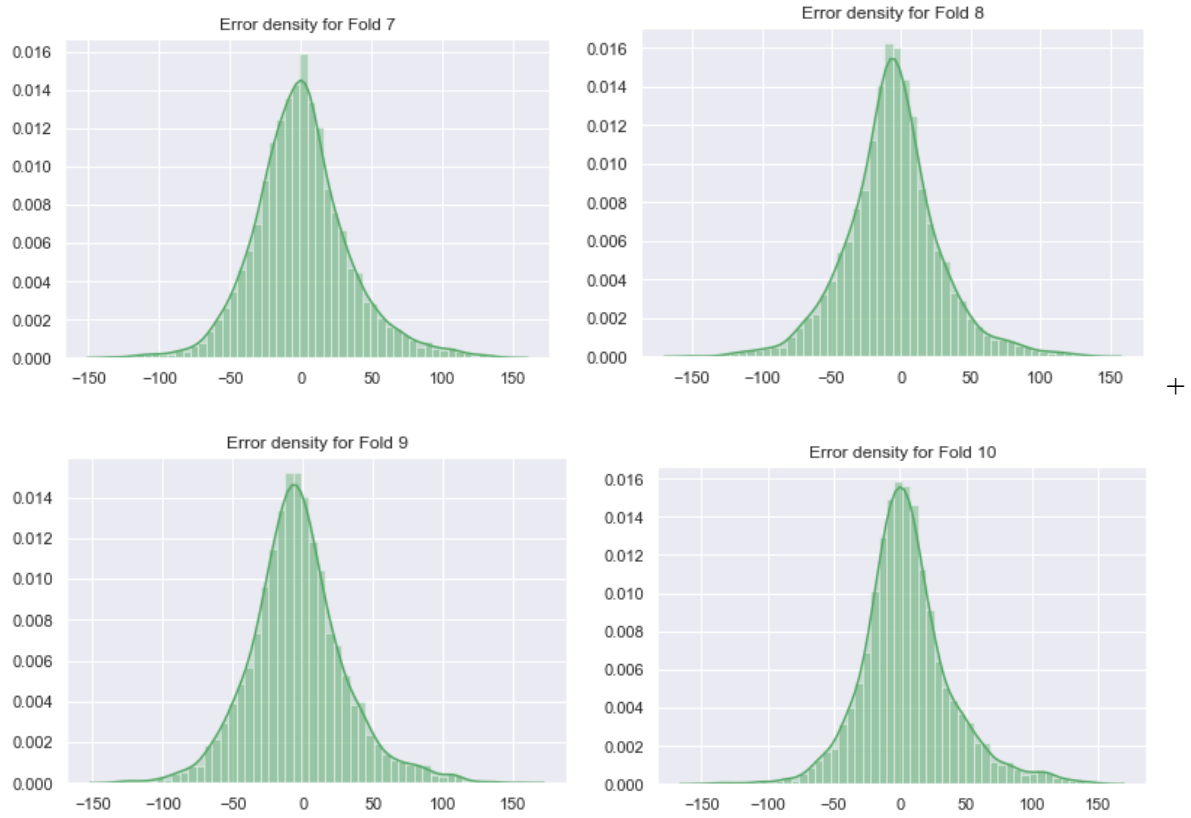
Accuracy = 0.88, Se = 0.84, Sp = 0.92, CC = 0.79, AUC = 0.94

Fig. 4. ROC-curves obtained in external validation of classification model.

Regression model

Fig.5 presents error density curves obtained in the 10-fold cross validation procedure for regression models. We splited the dataset into 10 subsets, each of unique subset had been selected as testing set, and training dataset constituted based on 9 remaining subsets. In all, we trained and evaluated 10 models The mean absolute error (MAE) for regression model was 25 nt.





Average MAE = 25 nt

Fig. 5. Error density curves obtained in 10-fold cross-validation of regression models.

We performed external validation of the model on the test chromosome. The mean absolute error (MAE) for the regression model was 28 nt.

Results and Discussion

We have developed an efficient, deep learning approach, for prediction of the position of transcription start sites in eukaryotes using nucleotide sequence. The approach is data-type independent and allows of incorporation of additional data types (such as RNA-seq and tissue specific DNA methylation), refining positions of TSS for tissue-specific and stress-specific expressions. We compared TransPrise with the CNNProm approach on an independent test set (chromosome 2) 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 TransPrize is more than twice larger than for CNNProm classification models (0.79 vs. 0.31), indicating the significantly higher efficiency of TransPrize 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 28 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 ($a_{i,j}$) - 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 of filter motifs correspond to known regulatory elements: TGGGCC (Lu et al. 2013), CGATT (Rose et al. 2016), ACTCAT(Chen et al. 2016), and CGCG box . Motif TGGGCC is targeted by the TCP transcription factor through its interaction with proliferating cell nuclear antigens PCF1 and PCF2(Lu et al. 2013); ACTCAT motif(Weltmeier et al. 2006) 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 (Yang & Poovaiah 2002).

Fig. 6 shows filter motifs that correspond to two well-characterized features of eukaryotic promoters: Initiator element CA and TATA-box(Troukhan et al. 2009)(Smale & Baltimore 1989)(Triska et al. 2017)(Zhu et al. 1995).

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 hypothesise that other features may correspond to unknown regulatory elements.

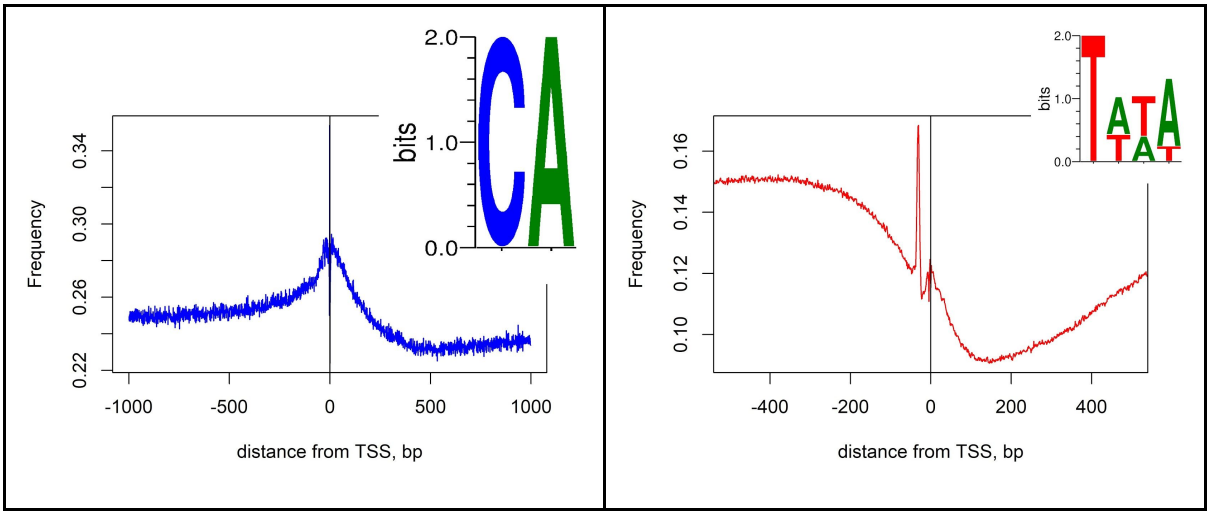


Fig 7. Distribution of CA (peaking at TSS) and TATA (peaking at TSS-30)

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 (https://hub.docker.com/r/zarubinaa/tss-rice/) with all necessary packages, models and code. The source code for our program TransPrise is available on GitHub (https://github.com/StepanAbstro/TransPrise) and is ready to use for new model training.

Bibliography

Alexandrov, N.N. et al., 2006. Features of Arabidopsis Genes and Genome Discovered using Full-length cDNAs. *Plant molecular biology*, 60(1), pp.69–85. Available at: http://dx.doi.org/10.1007/s11103-005-2564-9.

Alexandrov, N.N. et al., 2009. Insights into corn genes derived from large-scale cDNA sequencing. *Plant molecular biology*, 69(1-2), pp.179–194. Available at: http://dx.doi.org/10.1007/s11103-008-9415-4.

Atwell, S. et al., 2010. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, 465(7298), pp.627–631. Available at: http://dx.doi.org/10.1038/nature08800.

Batut, P. et al., 2013. High-fidelity promoter profiling reveals widespread alternative

promoter usage and transposon-driven developmental gene expression. *Genome research*, 23(1), pp.169–180. Available at: http://dx.doi.org/10.1101/gr.139618.112.

Berendzen, K.W. et al., 2006. Cis-motifs upstream of the transcription and translation initiation sites are effectively revealed by their positional disequilibrium in eukaryote genomes using frequency distribution curves. *BMC bioinformatics*, 7, p.522. Available at: http://dx.doi.org/10.1186/1471-2105-7-522.

Campbell, M.S. et al., 2014. Genome Annotation and Curation Using MAKER and MAKER-P. In *Current Protocols in Bioinformatics*. pp. 4.11.1–4.11.39. Available at: http://dx.doi.org/10.1002/0471250953.bi0411s48.

Carninci, P. et al., 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature genetics*, 38(6), pp.626–635. Available at: http://dx.doi.org/10.1038/ng1789.

Chan, K.-L. et al., 2017. Evidence-based gene models for structural and functional annotations of the oil palm genome. *Biology direct*, 12(1), p.21. Available at: http://dx.doi.org/10.1186/s13062-017-0191-4.

Chen, J. et al., 2016. ZmbZIP91 regulates expression of starch synthesis-related genes by binding to ACTCAT elements in their promoters. *Journal of experimental botany*, 67(5), pp.1327–1338. Available at: http://dx.doi.org/10.1093/jxb/erv527.

Davis, T.M. et al., 2010. An examination of targeted gene neighborhoods in strawberry. *BMC plant biology*, 10, p.81. Available at: http://dx.doi.org/10.1186/1471-2229-10-81.

Eckardt, N.A., 2014. Unexpected Structure of Plant Promoters: *The Plant cell*, 26(7), pp.2726–2726. Available at: http://dx.doi.org/10.1105/tpc.114.129239.

Elhaik, E., Pellegrini, M. & Tatarinova, T.V., 2014. Gene expression and nucleotide composition are associated with genic methylation level in Oryza sativa. *BMC bioinformatics*, 15, p.23. Available at: http://dx.doi.org/10.1186/1471-2105-15-23.

Elhaik, E. & Tatarinova, T., 2012. *GC3 Biology in Eukaryotes and Prokaryotes*, Available at: https://books.google.com/books/about/GC3_Biology_in_Eukaryotes_and_Prokaryote.html?hl=&id=JYTzoAEACAAJ.

Farrell, R.E. & Bassett, C.L., Multiple Transcript Initiation as a Mechanism for Regulating Gene Expression. In *Regulation of Gene Expression in Plants*. pp. 39–66. Available at: http://dx.doi.org/10.1007/978-0-387-35640-2_2.

Herbig, A., Sharma, C. & Nieselt, K., 2013. Automated transcription start site prediction for comparative Transcriptomics using the SuperGenome. *EMBnet.journal*, 19(A), p.19. Available at: http://dx.doi.org/10.14806/ej.19.a.617.

Hoff, K.J. et al., 2015. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1. *Bioinformatics* , 32(5), pp.767–769. Available at: http://dx.doi.org/10.1093/bioinformatics/btv661.

Holt, C. & Yandell, M., 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC bioinformatics*, 12(1), p.491. Available at: http://dx.doi.org/10.1186/1471-2105-12-491.

Ito, Y. et al., 2005. Rice Annotation Database (RAD): a contig-oriented database for map-based rice genomics. *Nucleic acids research*, 33(Database issue), pp.D651–5. Available at: http://dx.doi.org/10.1093/nar/gki083.

Jiang, Y. et al., 2015. Genetic fine mapping and candidate gene analysis of the Gossypium hirsutum Ligon lintless-1 (Li1) mutant on chromosome 22(D). *Molecular genetics and genomics: MGG*, 290(6), pp.2199–2211. Available at: http://dx.doi.org/10.1007/s00438-015-1070-2.

Kawahara, Y. et al., 2013. Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice* , 6(1), p.4. Available at: http://dx.doi.org/10.1186/1939-8433-6-4.

Kawaji, H. et al., 2006. CAGE Basic/Analysis Databases: the CAGE resource for comprehensive promoter analysis. *Nucleic acids research*, 34(Database issue), pp.D632–6. Available at: http://dx.doi.org/10.1093/nar/gkj034.

Kawaji, H. et al., 2014. Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome research*, 24(4), pp.708–717. Available at: http://dx.doi.org/10.1101/gr.156232.113.

Keilwagen, J. et al., 2018. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC bioinformatics*, 19(1), p.189. Available at: http://dx.doi.org/10.1186/s12859-018-2203-5.

Li, M.J. et al., 2015. Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression. *Briefings in bioinformatics*, 16(3), pp.393–412. Available at: http://dx.doi.org/10.1093/bib/bbu018.

Liseron-Monfils, C. et al., 2013. Promzea: a pipeline for discovery of co-regulatory motifs in maize and other plant species and its application to the anthocyanin and phlobaphene biosynthetic pathways and the Maize Development Atlas. *BMC plant biology*, 13, p.42. Available at: http://dx.doi.org/10.1186/1471-2229-13-42.

Louzada, E.S., Alternative Processing as a Mechanism for Regulating Gene Expression. In *Regulation of Gene Expression in Plants*. pp. 67–100. Available at: http://dx.doi.org/10.1007/978-0-387-35640-2_3.

Lukashin, A., 1998. GeneMark.hmm: new solutions for gene finding. *Nucleic acids research*, 26(4), pp.1107–1115. Available at: http://dx.doi.org/10.1093/nar/26.4.1107.

Lu, Z. et al., 2013. Genome-wide binding analysis of the transcription activator ideal plant architecture1 reveals a complex network regulating rice plant architecture. *The Plant cell*, 25(10), pp.3743–3759. Available at: http://dx.doi.org/10.1105/tpc.113.113639.

Morton, T. et al., 2014. Paired-end analysis of transcription start sites in Arabidopsis reveals plant-specific promoter signatures. *The Plant cell*, 26(7), pp.2746–2760. Available at: http://dx.doi.org/10.1105/tpc.114.125617.

Mulder, N., 2018. Faculty of 1000 evaluation for Identifying noncoding risk variants using disease-relevant gene regulatory networks. *F1000 - Post-publication peer review of the biomedical literature*. Available at: http://dx.doi.org/10.3410/f.732691059.793543018.

Nasiri, J. et al., 2013. Gene identification programs in bread wheat: a comparison study. *Nucleosides, nucleotides & nucleic acids*, 32(10), pp.529–554. Available at: http://dx.doi.org/10.1080/15257770.2013.832773.

Pritsker, M. et al., 2004. Whole-genome discovery of transcription factor binding sites by network-level conservation. *Genome research*, 14(1), pp.99–108. Available at: http://dx.doi.org/10.1101/gr.1739204.

Rose, A.B. et al., 2016. Intron sequences that stimulate gene expression in Arabidopsis. *Plant molecular biology*, 92(3), pp.337–346. Available at: http://dx.doi.org/10.1007/s11103-016-0516-1.

Rye, M. et al., 2014. Chromatin states reveal functional associations for globally defined transcription start sites in four human cell lines. *BMC genomics*, 15, p.120. Available at: http://dx.doi.org/10.1186/1471-2164-15-120.

Salamov, A.A. & Solovyev, V.V., 2000. Ab initio gene finding in Drosophila genomic DNA. *Genome research*, 10(4), pp.516–522. Available at: https://www.ncbi.nlm.nih.gov/pubmed/10779491.

Sandelin, A. et al., 2007. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature reviews. Genetics*, 8(6), pp.424–436. Available at: http://dx.doi.org/10.1038/nrg2026.

Sanusi, N.S.N.M. et al., 2018. PalmXplore: oil palm gene database. *Database: the journal of biological databases and curation*, 2018. Available at: http://dx.doi.org/10.1093/database/bay095.

Shahmuradov, I.A. & Solovyev, V.V., 2015. Nsite, NsiteH and NsiteM computer tools for studying transcription regulatory elements: Fig. 1. *Bioinformatics* , 31(21), pp.3544–3545. Available at: http://dx.doi.org/10.1093/bioinformatics/btv404.

Shahmuradov, I.A., Umarov, R.K. & Solovyev, V.V., 2017. TSSPlant: a new tool for prediction of plant Pol II promoters. *Nucleic acids research*, 45(8), p.e65. Available at: http://dx.doi.org/10.1093/nar/gkw1353.

Sheshadri, S.A. et al., 2018. Comparative genome based cis-elements analysis in the 5’ upstream and 3' downstream region of cell wall invertase and Phenylalanine ammonia lyase in Nicotiana benthamiana. *Computational biology and chemistry*, 72, pp.181–191. Available at: http://dx.doi.org/10.1016/j.compbiolchem.2017.11.004.

Smale, S.T. & Baltimore, D., 1989. The “initiator” as a transcription control element. *Cell*,

57(1), pp.103–113. Available at: http://dx.doi.org/10.1016/0092-8674(89)90176-1.

Solovyev, V., 2003. PromH: promoters identification using orthologous genomic sequences. *Nucleic acids research*, 31(13), pp.3540–3545. Available at: http://dx.doi.org/10.1093/nar/gkg525.

Solovyev, V.V. & Shahmuradov, I.A., 2003. PromH: Promoters identification using orthologous genomic sequences. *Nucleic acids research*, 31(13), pp.3540–3545. Available at: https://www.ncbi.nlm.nih.gov/pubmed/12824362.

Solovyev, V.V., Shahmuradov, I.A. & Salamov, A.A., 2010. Identification of Promoter Regions and Regulatory Sites. In *Methods in Molecular Biology*. pp. 57–83. Available at: http://dx.doi.org/10.1007/978-1-60761-854-6_5.

Stanke, M. & Morgenstern, B., 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic acids research*, 33(Web Server issue), pp.W465–7. Available at: http://dx.doi.org/10.1093/nar/gki458.

Steijger, T. et al., 2013. Assessment of transcript reconstruction methods for RNA-seq. *Nature methods*, 10(12), pp.1177–1184. Available at: http://dx.doi.org/10.1038/nmeth.2714.

Tatarinova, T., Kryshchenko, A., et al., 2013. NPEST: a nonparametric method and a database for transcription start site prediction. *Quantitative Biology*, 1(4), pp.261–271. Available at: http://dx.doi.org/10.1007/s40484-013-0022-2.

Tatarinova, T., Elhaik, E. & Pellegrini, M., 2013. Cross-species analysis of genic GC3 content and DNA methylation patterns. *Genome biology and evolution*, 5(8), pp.1443–1456. Available at: http://dx.doi.org/10.1093/gbe/evt103.

Tatarinova, T.V. et al., 2016. Nucleotide diversity analysis highlights functionally important genomic regions. *Scientific reports*, 6, p.35730. Available at: http://dx.doi.org/10.1038/srep35730.

Triska, M. et al., 2017. Nucleotide patterns aiding in prediction of eukaryotic promoters. *PloS one*, 12(11), p.e0187243. Available at: http://dx.doi.org/10.1371/journal.pone.0187243.

Troukhan, M. et al., 2009. Genome-Wide Discovery of cis-Elements in Promoter Sequences Using Gene Expression. *Omics: a journal of integrative biology*, 13(2), pp.139–151. Available at: http://dx.doi.org/10.1089/omi.2008.0034.

Umarov, R.K. & Solovyev, V.V., 2017. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PloS one*, 12(2), p.e0171410. Available at: http://dx.doi.org/10.1371/journal.pone.0171410.

Weltmeier, F. et al., 2006. Combinatorial control of Arabidopsis proline dehydrogenase transcription by specific heterodimerisation of bZIP transcription factors. *The EMBO journal*, 25(13), pp.3133–3143. Available at: http://dx.doi.org/10.1038/sj.emboj.7601206.

Yang, T. & Poovaiah, B.W., 2002. A calmodulin-binding/CGCG box DNA-binding protein family involved in multiple signaling pathways in plants. *The Journal of biological chemistry*, 277(47), pp.45049–45058. Available at: http://dx.doi.org/10.1074/jbc.M207941200.

Yao, H. et al., 2005. Evaluation of five ab initio gene prediction programs for the discovery of maize genes. *Plant molecular biology*, 57(3), pp.445–460. Available at: http://dx.doi.org/10.1007/s11103-005-0271-1.

Zhang, S.-L. et al., 2008. The Prediction of Rice Gene by Fgenesh. *Agricultural sciences in China / sponsored by the Chinese Academy of Agricultural Sciences*, 7(4), pp.387–394. Available at: http://dx.doi.org/10.1016/s1671-2927(08)60081-4.

Zhu, Q., Dabi, T. & Lamb, C., 1995. TATA Box and Initiator Functions in the Accurate Transcription of a Plant Minimal Promoter in vitro. *The Plant cell*, 7(10), p.1681. Available at: http://dx.doi.org/10.2307/3870029.

Alexandrov, N.N. et al., 2006. Features of Arabidopsis Genes and Genome Discovered using Full-length cDNAs. Plant molecular biology, 60(1), pp.69–85. Available at: http://dx.doi.org/10.1007/s11103-005-2564-9.

Alexandrov, N.N. et al., 2009. Insights into corn genes derived from large-scale cDNA sequencing. Plant molecular biology, 69(1-2), pp.179–194. Available at: http://dx.doi.org/10.1007/s11103-008-9415-4.

Atwell, S. et al., 2010. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature, 465(7298), pp.627–631. Available at: http://dx.doi.org/10.1038/nature08800.

Batut, P. et al., 2013. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. Genome research, 23(1), pp.169–180. Available at: http://dx.doi.org/10.1101/gr.139618.112.

Berendzen, K.W. et al., 2006. Cis-motifs upstream of the transcription and translation initiation sites are effectively revealed by their positional disequilibrium in eukaryote genomes using frequency distribution curves. BMC bioinformatics, 7, p.522. Available at: http://dx.doi.org/10.1186/1471-2105-7-522.

Campbell, M.S. et al., 2014. Genome Annotation and Curation Using MAKER and MAKER-P. In Current Protocols in Bioinformatics. pp. 4.11.1–4.11.39. Available at: http://dx.doi.org/10.1002/0471250953.bi0411s48.

Carninci, P. et al., 2006. Genome-wide analysis of mammalian promoter architecture and evolution. Nature genetics, 38(6), pp.626–635. Available at: http://dx.doi.org/10.1038/ng1789.

Chan, K.-L. et al., 2017. Evidence-based gene models for structural and functional annotations of the oil palm genome. Biology direct, 12(1), p.21. Available at: http://dx.doi.org/10.1186/s13062-017-0191-4.

Chen, J. et al., 2016. ZmbZIP91 regulates expression of starch synthesis-related genes by binding to ACTCAT elements in their promoters. Journal of experimental botany, 67(5), pp.1327–1338. Available at: http://dx.doi.org/10.1093/jxb/erv527.

Davis, T.M. et al., 2010. An examination of targeted gene neighborhoods in strawberry. BMC plant biology, 10, p.81. Available at: http://dx.doi.org/10.1186/1471-2229-10-81.

Eckardt, N.A., 2014. Unexpected Structure of Plant Promoters: The Plant cell, 26(7), pp.2726–2726. Available at: http://dx.doi.org/10.1105/tpc.114.129239.

Elhaik, E., Pellegrini, M. & Tatarinova, T.V., 2014. Gene expression and nucleotide composition are associated with genic methylation level in Oryza sativa. BMC bioinformatics, 15, p.23. Available at: http://dx.doi.org/10.1186/1471-2105-15-23.

Elhaik, E. & Tatarinova, T., 2012. GC3 Biology in Eukaryotes and Prokaryotes, Available at: https://books.google.com/books/about/GC3_Biology_in_Eukaryotes_and_Prokaryote.html?hl=&id=JYTzoAEACAAJ.

Farrell, R.E. & Bassett, C.L., Multiple Transcript Initiation as a Mechanism for Regulating Gene Expression. In Regulation of Gene Expression in Plants. pp. 39–66. Available at: http://dx.doi.org/10.1007/978-0-387-35640-2_2.

Herbig, A., Sharma, C. & Nieselt, K., 2013. Automated transcription start site prediction for comparative Transcriptomics using the SuperGenome. EMBnet.journal, 19(A), p.19. Available at: http://dx.doi.org/10.14806/ej.19.a.617.

Hoff, K.J. et al., 2015. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1. Bioinformatics , 32(5), pp.767–769. Available at: http://dx.doi.org/10.1093/bioinformatics/btv661.

Holt, C. & Yandell, M., 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC bioinformatics, 12(1), p.491. Available at: http://dx.doi.org/10.1186/1471-2105-12-491.

Ito, Y. et al., 2005. Rice Annotation Database (RAD): a contig-oriented database for map-based rice genomics. Nucleic acids research, 33(Database issue), pp.D651–5. Available at: http://dx.doi.org/10.1093/nar/gki083.

Jiang, Y. et al., 2015. Genetic fine mapping and candidate gene analysis of the Gossypium hirsutum Ligon lintless-1 (Li1) mutant on chromosome 22(D). Molecular genetics and genomics: MGG, 290(6), pp.2199–2211. Available at: http://dx.doi.org/10.1007/s00438-015-1070-2.

Kawahara, Y. et al., 2013. Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice , 6(1), p.4. Available at: http://dx.doi.org/10.1186/1939-8433-6-4.

Kawaji, H. et al., 2006. CAGE Basic/Analysis Databases: the CAGE resource for comprehensive promoter analysis. Nucleic acids research, 34(Database issue),

pp.D632–6. Available at: http://dx.doi.org/10.1093/nar/gkj034.

Kawaji, H. et al., 2014. Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. Genome research, 24(4), pp.708–717. Available at: http://dx.doi.org/10.1101/gr.156232.113.

Keilwagen, J. et al., 2018. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. BMC bioinformatics, 19(1), p.189. Available at: http://dx.doi.org/10.1186/s12859-018-2203-5.

Li, M.J. et al., 2015. Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression. Briefings in bioinformatics, 16(3), pp.393–412. Available at: http://dx.doi.org/10.1093/bib/bbu018.

Liseron-Monfils, C. et al., 2013. Promzea: a pipeline for discovery of co-regulatory motifs in maize and other plant species and its application to the anthocyanin and phlobaphene biosynthetic pathways and the Maize Development Atlas. BMC plant biology, 13, p.42. Available at: http://dx.doi.org/10.1186/1471-2229-13-42.

Louzada, E.S., Alternative Processing as a Mechanism for Regulating Gene Expression. In Regulation of Gene Expression in Plants. pp. 67–100. Available at: http://dx.doi.org/10.1007/978-0-387-35640-2_3.

Lukashin, A., 1998. GeneMark.hmm: new solutions for gene finding. Nucleic acids research, 26(4), pp.1107–1115. Available at: http://dx.doi.org/10.1093/nar/26.4.1107.

Lu, Z. et al., 2013. Genome-wide binding analysis of the transcription activator ideal plant architecture1 reveals a complex network regulating rice plant architecture. The Plant cell, 25(10), pp.3743–3759. Available at: http://dx.doi.org/10.1105/tpc.113.113639.

Morton, T. et al., 2014. Paired-end analysis of transcription start sites in Arabidopsis reveals plant-specific promoter signatures. The Plant cell, 26(7), pp.2746–2760. Available at: http://dx.doi.org/10.1105/tpc.114.125617.

Mulder, N., 2018. Faculty of 1000 evaluation for Identifying noncoding risk variants using disease-relevant gene regulatory networks. F1000 - Post-publication peer review of the biomedical literature. Available at: http://dx.doi.org/10.3410/f.732691059.793543018.

Nasiri, J. et al., 2013. Gene identification programs in bread wheat: a comparison study. Nucleosides, nucleotides & nucleic acids, 32(10), pp.529–554. Available at: http://dx.doi.org/10.1080/15257770.2013.832773.

Pritsker, M. et al., 2004. Whole-genome discovery of transcription factor binding sites by network-level conservation. Genome research, 14(1), pp.99–108. Available at: http://dx.doi.org/10.1101/gr.1739204.

Rose, A.B. et al., 2016. Intron sequences that stimulate gene expression in Arabidopsis. Plant molecular biology, 92(3), pp.337–346. Available at:

http://dx.doi.org/10.1007/s11103-016-0516-1.

Rye, M. et al., 2014. Chromatin states reveal functional associations for globally defined transcription start sites in four human cell lines. BMC genomics, 15, p.120. Available at: http://dx.doi.org/10.1186/1471-2164-15-120.

Salamov, A.A. & Solovyev, V.V., 2000. Ab initio gene finding in Drosophila genomic DNA. Genome research, 10(4), pp.516–522. Available at: https://www.ncbi.nlm.nih.gov/pubmed/10779491.

Sandelin, A. et al., 2007. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nature reviews. Genetics, 8(6), pp.424–436. Available at: http://dx.doi.org/10.1038/nrg2026.

Sanusi, N.S.N.M. et al., 2018. PalmXplore: oil palm gene database. Database: the journal of biological databases and curation, 2018. Available at: http://dx.doi.org/10.1093/database/bay095.

Shahmuradov, I.A. & Solovyev, V.V., 2015. Nsite, NsiteH and NsiteM computer tools for studying transcription regulatory elements: Fig. 1. Bioinformatics , 31(21), pp.3544–3545. Available at: http://dx.doi.org/10.1093/bioinformatics/btv404.

Shahmuradov, I.A., Umarov, R.K. & Solovyev, V.V., 2017. TSSPlant: a new tool for prediction of plant Pol II promoters. Nucleic acids research, 45(8), p.e65. Available at: http://dx.doi.org/10.1093/nar/gkw1353.

Sheshadri, S.A. et al., 2018. Comparative genome based cis-elements analysis in the 5' upstream and 3' downstream region of cell wall invertase and Phenylalanine ammonia lyase in Nicotiana benthamiana. Computational biology and chemistry, 72, pp.181–191. Available at: http://dx.doi.org/10.1016/j.compbiolchem.2017.11.004.

Smale, S.T. & Baltimore, D., 1989. The "initiator" as a transcription control element. Cell, 57(1), pp.103–113. Available at: http://dx.doi.org/10.1016/0092-8674(89)90176-1.

Solovyev, V., 2003. PromH: promoters identification using orthologous genomic sequences. Nucleic acids research, 31(13), pp.3540–3545. Available at: http://dx.doi.org/10.1093/nar/gkg525.

Solovyev, V.V. & Shahmuradov, I.A., 2003. PromH: Promoters identification using orthologous genomic sequences. Nucleic acids research, 31(13), pp.3540–3545. Available at: https://www.ncbi.nlm.nih.gov/pubmed/12824362.

Solovyev, V.V., Shahmuradov, I.A. & Salamov, A.A., 2010. Identification of Promoter Regions and Regulatory Sites. In Methods in Molecular Biology. pp. 57–83. Available at: http://dx.doi.org/10.1007/978-1-60761-854-6_5.

Stanke, M. & Morgenstern, B., 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic acids research, 33(Web Server issue), pp.W465–7. Available at: http://dx.doi.org/10.1093/nar/gki458.

Steijger, T. et al., 2013. Assessment of transcript reconstruction methods for RNA-seq. Nature methods, 10(12), pp.1177–1184. Available at: http://dx.doi.org/10.1038/nmeth.2714.

Tatarinova, T., Kryshchenko, A., et al., 2013. NPEST: a nonparametric method and a database for transcription start site prediction. Quantitative Biology, 1(4), pp.261–271. Available at: http://dx.doi.org/10.1007/s40484-013-0022-2.

Tatarinova, T., Elhaik, E. & Pellegrini, M., 2013. Cross-species analysis of genic GC3 content and DNA methylation patterns. Genome biology and evolution, 5(8), pp.1443–1456. Available at: http://dx.doi.org/10.1093/gbe/evt103.

Tatarinova, T.V. et al., 2016. Nucleotide diversity analysis highlights functionally important genomic regions. Scientific reports, 6, p.35730. Available at: http://dx.doi.org/10.1038/srep35730.

Triska, M. et al., 2017. Nucleotide patterns aiding in prediction of eukaryotic promoters. PloS one, 12(11), p.e0187243. Available at: http://dx.doi.org/10.1371/journal.pone.0187243.

Troukhan, M. et al., 2009. Genome-Wide Discovery of cis-Elements in Promoter Sequences Using Gene Expression. Omics: a journal of integrative biology, 13(2), pp.139–151. Available at: http://dx.doi.org/10.1089/omi.2008.0034.

Umarov, R.K. & Solovyev, V.V., 2017. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. PloS one, 12(2), p.e0171410. Available at: http://dx.doi.org/10.1371/journal.pone.0171410.

Weltmeier, F. et al., 2006. Combinatorial control of Arabidopsis proline dehydrogenase transcription by specific heterodimerisation of bZIP transcription factors. The EMBO journal, 25(13), pp.3133–3143. Available at: http://dx.doi.org/10.1038/sj.emboj.7601206.

Yang, T. & Poovaiah, B.W., 2002. A calmodulin-binding/CGCG box DNA-binding protein family involved in multiple signaling pathways in plants. The Journal of biological chemistry, 277(47), pp.45049–45058. Available at: http://dx.doi.org/10.1074/jbc.M207941200.

Yao, H. et al., 2005. Evaluation of five ab initio gene prediction programs for the discovery of maize genes. Plant molecular biology, 57(3), pp.445–460. Available at: http://dx.doi.org/10.1007/s11103-005-0271-1.

Zhang, S.-L. et al., 2008. The Prediction of Rice Gene by Fgenesh. Agricultural sciences in China / sponsored by the Chinese Academy of Agricultural Sciences, 7(4), pp.387–394. Available at: http://dx.doi.org/10.1016/s1671-2927(08)60081-4.

Zhu, Q., Dabi, T. & Lamb, C., 1995. TATA Box and Initiator Functions in the Accurate Transcription of a Plant Minimal Promoter in vitro. The Plant cell, 7(10), p.1681. Available at: http://dx.doi.org/10.2307/3870029.