Sequence analysis

Promoter analysis and prediction in the human genome using sequence-based deep learning models

Ramzan Umarov [1], Hiroyuki Kuwahara [1], Yu Li [1], Xin Gao [1,*] and Victor Solovyev [2,*]

[1] King Abdullah University of Science and Technology, Computational Bioscience Research Center, Computer, Electrical and Mathematical Sciences and Engineering Division, Thuwal 23955-6900, Saudi Arabia and
[2] Institute of Cytology and Genetics SB RAS, 10 Lavrentiev Ave., Novosibirsk 630090, Russia.

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Computational identification of promoters is notoriously difficult as human genes often have unique promoter sequences that provide regulation of transcription and interaction with transcription initiation complex. While there are many attempts to develop computational promoter identification methods, we have no reliable tool to analyze long genomic sequences.
Results: In this work we further develop our deep learning approach that was relatively successful to discriminate short promoter and non-promoter sequences. Instead of focusing on the classification accuracy, in this work we predict the exact positions of the TSS inside the genomic sequences testing every possible location. We studied human promoters to find effective regions for discrimination and built corresponding deep learning models. These models use adaptively constructed negative set, which iteratively improves the model's discriminative ability. Our method significantly outperforms the previously developed promoter prediction programs by considerably reducing the number of false positive predictions. We have achieved error-per-1000-bp rate of 0.02 and have 0.31 errors per correct prediction, which is significantly better than the results of other human promoter predictors.
Availability: The developed method is available as a web server at http://www.cbrc.kaust.edu.sa/PromID/.
Contact: Victor Solovyev Email: solovictor@gmail.com; Xin Gao Email:xin.gao@kaust.edu.sa.

1 Introduction

The high fidelity of the RNA polymerase II (pol II) transcription system is necessary for precise spatiotemporal regulation of endogenous protein expression and essential to proper development and homeostasis in eukaryotes. Among the key *cis*-regulatory modules for RNA pol II-mediated transcription is the core promoter, which is typically situated within a DNA segment spanning from -40 bp to +40 bp relative to the transcription start site (TSS) at position +1 (Kadonaga, 2012; Danino *et al.*, 2015; Vo Ngoc *et al.*, 2017a). This stretch of DNA serves as a platform on which RNA pol II and a number of auxiliary factors assemble into the transcription machinery, which is capable of integrating a range of intrinsic and extrinsic signals, to ultimately determine the proper initiation of transcription (Lodish *et al.*, 2000; Butler and Kadonaga, 2002; Morris *et al.*, 2004; Juven-Gershon *et al.*, 2008; Kadonaga, 2012; Roy and Singer, 2015; Zabidi *et al.*, 2015; Vo Ngoc *et al.*, 2017b). Thus, the characterization of the structure-function relation of the core promoter is crucial to unraveling the complex molecular control mechanisms underlying not just the constitutive basal expression but also the regulated expression in the RNA pol II transcription system.

Decades of *in vitro* research has identified a number of functional sequence motifs for the RNA pol II core promoter (Butler and Kadonaga, 2002; Smale and Kadonaga, 2003; Roy and Singer, 2015; Vo Ngoc *et al.*, 2017b). Among such functional core promoter elements, perhaps, the

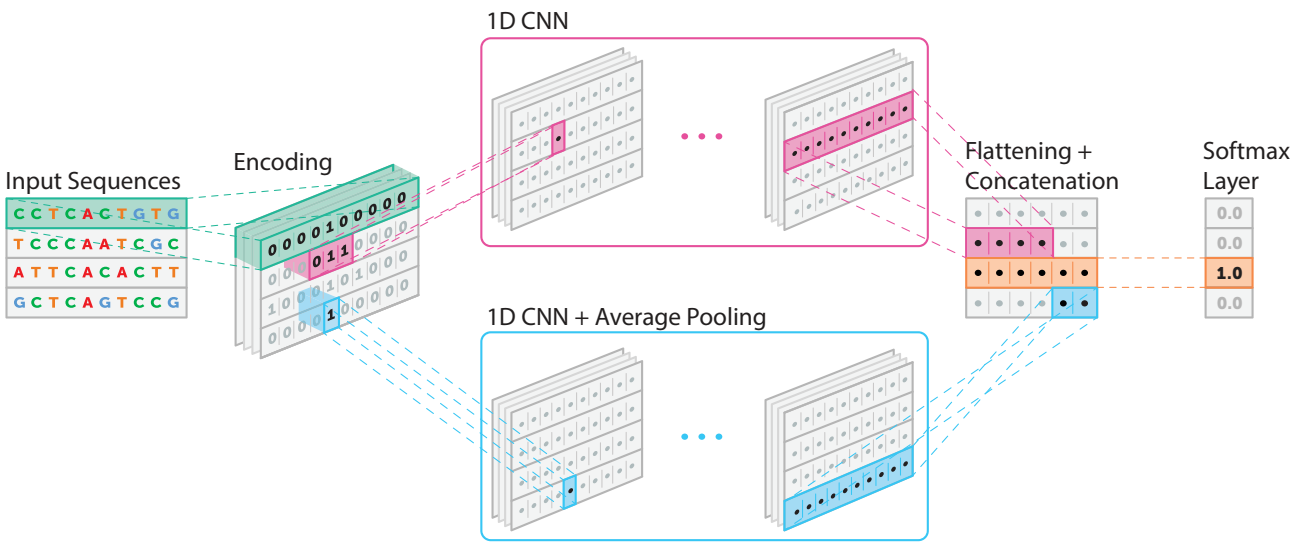


Fig. 1. Deep learning model architecture that was used in building promoter models of DeeReCT-PromID (see text for its description).

most well-known is the TATA-box, which was, in the past, thought to be universally present in RNA pol II core promoters (Vo Ngoc *et al.*, 2017b). However, the advent in genome-wide TSS detections based on high-throughput sequencing revealed that the core promoter structure is highly diverse and complex, and there are no universal core promoter elements (Lenhard *et al.*, 2012; Kadonaga, 2012; Roy and Singer, 2015; Zabidi *et al.*, 2015; Arnold *et al.*, 2017; Vo Ngoc *et al.*, 2017b). Indeed, recent estimates showed that only about 17% of eukaryotic core promoters contain the TATA-box (Yella and Bansal, 2017). More surprisingly, genome-wide structural analysis found that many core promoters do not possess any of the known core promoter elements. Such structural heterogeneity permits the core promoter to expand its functional repertories so as to serve as gene- and cell-type-specific transcription regulator that responds to a range of conditions; however, because of this large diversity, the design principle of the core promoter still remains largely elusive (Roy and Singer, 2015; Arnold *et al.*, 2017; Garieri *et al.*, 2017; Vo Ngoc *et al.*, 2017b).

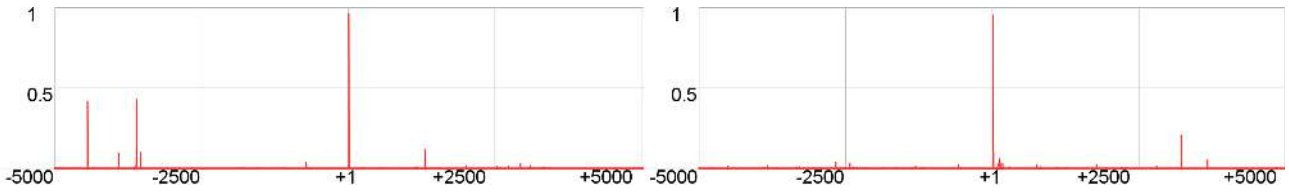
The structure of the human promoter is notoriously complex and diverse. One explanation for this is that such complex and diverse structures must be "designed" to properly control expression of ∼25,000 protein coding genes based on interactions with only ∼1,850 transcription factors in the human genome (Maston *et al.*, 2006). Another explanation comes from a molecular evolution study which discovered substantially accelerated rates of evolution in primate promoters compared with other mammalian promoters (Taylor *et al.*, 2006). This rapid primate promoter evolution was found to be comparable to the neutral substitution rate, suggesting that primate promoters have weak selective constraints, and this suggestion can also explain highly complex and diverse structures in the human promoter. In any case, a better understanding of the structure-function relation of the human promoter has particularly important implications as some genetic variants in such noncoding regions are associated with rare Mendelian diseases (Edwards *et al.*, 2013; Rojano *et al.*, 2018). Furthermore, some cancer cells are associated with somatic mutations in promoter regions (Vinagre *et al.*, 2013; Fredriksson *et al.*, 2017). In order to gain insights into what types of genetic variations can cause aberrant expression leading to human diseases, it is crucial to accurately predict the locations of human promoters and to understand their structural patterns.

Here, we introduce DeeReCT-PromID, a novel machine learning-based approach for the prediction of human RNA pol II core promoters. Taking advantage of the big promoter collection with experimentally validated TSSs (Dreos *et al.*, 2016) generated by modern high-throughput techniques, we build a deep learning model using sequence data as an input. To avoid bias based on prior knowledge about promoter loci (e.g, sequences with known core promoter elements and high density of CpG dinucleotides), we do not use predefined features; but rather attempt to discover sequence features and learn salient patterns of the human promoter solely from the training set. This is important especially in the prediction of human promoters since the structural features of many promoters are still unknown (Maston *et al.*, 2006; Roy and Singer, 2015). We previously developed a similar convolutional neural network-based algorithm for the prediction of core promoter locations in several model organisms (Umarov and Solovyev, 2017). While this method was able to outperform previously developed promoter prediction methods (Umarov and Solovyev, 2017), its false positive rate was not adequate enough to ensure the accurate detection of promoters on long genomic sequences. DeeReCT-PromID was developed to chiefly alleviate this limitation and to focus on the promoter prediction on longer sequences. Specifically, to reduce the false positive rate, we adaptively and iteratively train the predictor by changing the distribution of samples in the training set based on the false-positive errors it made in the previous iteration. By including difficult non-promoter sequences in the training set, we can force the predictor to learn promoter patterns to rule out such sequences. To evaluate the performance of the new method, we compared our method with publicly available tools for the human promoter prediction task. We found that DeeReCT-PromID outperformed the other predictors and achieved a much smaller error-per-1000-bp rate. Our results demonstrate the usefulness of the proposed method for the human promoter prediction on long genomic sequences and suggest its potential value as a tool to gain insights into the design principle for the human core promoters.

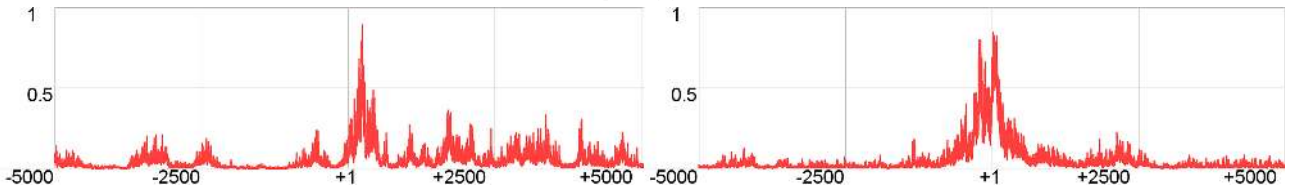
2 MATERIALS AND METHODS

2.1 Datasets

Our models are trained using human promoter sequences extracted from the EPDnew database (Dreos *et al.*, 2016). The EPD database is an annotated non-redundant collection of eukaryotic POL II promoters, for which the transcription start site has been determined experimentally. The



(a) Promoters with the TATA-box. Corresponding genes from left are COCH_1 and CCL5_1.



(b) Promoters without the TATA-box. Corresponding genes from left are FAM135A_1 and ASCC3_1.

Fig. 2. Scoring landscapes constructed by our models. True TSS is at position +1.

authors of the EPDnew database have demonstrated its higher quality over the ENSEMBL-derived (Aken *et al.*, 2016) human promoter set (Dreos *et al.*, 2012).

In this study we downloaded 16455 genomic sequences (from -5000 bp to +5000 bp, where +1 is a TSS position) containing human promoters from the EPD database. We used 90% of the sequences for training and 10% for testing. Positive and negative sets were extracted from the training set. A promoter region of a given size around the known TSS is considered to be a positive sequence. A negative sequence is the one outside the promoter region, which does not contain a known TSS. Initially, the negative set had the same size as the positive one and consisted of randomly picked negative sequences.

2.2 Deep neural network model

We use deep neural networks to identify promoter regions. The data is read in the fasta format and then transformed using one hot encoding. This encoding uses a vector of size 4 to represent each nucleotide. *A* is encoded as (1 0 0 0), *T* is encoded as (0 1 0 0), *G* is encoded as (0 0 1 0), and *C* is encoded as (0 0 0 1).

Our architecture consists of two Convolutional Neural Networks (CNN) which are in parallel (Figure 1). This means that they both have access to the original input. One CNN has an average pooling layer, while the other does not have any pooling layer. CNN with an average pooling layer has filter length one, while the other one uses filter length 15 and both have two convolutional layers. Average pooling is well suited to capture GC content of the sequence, which is known to be higher in a promoter region (Fenouil *et al.*, 2012), since we only care about the count of *G* and *C* nucleotides, not their positions inside a promoter. However, usage of a pooling layer deteriorates positional information which is important for some promoter elements that are located at a specific location in the promoter region, for example the initiator. Using two CNNs in parallel, we solved this conflict and were able to capture various significant promoter features. The outputs of the CNNs are concatenated, flattened and fed into a softmax layer which predicts the probability that an input sequence is a promoter.

Weight decay and dropout (Srivastava *et al.*, 2014) are used to improve the generalization capability of our model. Weight decay effectively limits the number of free parameters in the model to avoid overfitting. Introducing weight decay makes it possible to regularize the cost function by penalizing large weights. The main idea of dropout is to randomly set some nodes of the neural network to zero during training to prevent co-dependency among them. During the training we use dropout for the feature vector with keep probability of 0.5. Adam optimization algorithm is used to train the weights (Kingma and Ba, 2014), which is an improved version of stochastic gradient descent. We use TensorFlow (Abadi *et al.*, 2016) as the framework to construct the deep neural network. The training was performed on a workstation with two 980 GTX GPUs and took on average 7 hours.

2.3 Classification procedure

We train the model using positive and negative sets which consist of relatively short sequences with a fixed length. As our model accepts sequences of certain length as input we apply a sliding window approach to analyze long genomic sequences. This window is moved across the sequence and at each position the subsequence is fed into our model. The model gives us a score from 0 to 1 that represents the likelihood that a subsequence is a promoter region. If we plot these promoter scores, we will receive a scoring landscape of our model, see Figure 2.

If the value of the score of a sliding window is above the threshold it is predicted as a start of a promoter region. In practice, we construct two deep learning models - one is for identification of promoter sequences with the TATA-box and one for promoters without it. Promoters can be predicted much more accurately if they have the TATA-box, that is why we firstly apply the model trained specifically for the promoters with the TATA-box (TATA+ model). Next, we apply the model trained with the promoter sequences without the TATA-box (TATA- model). We account the second model predictions that are not too close to the first model predictions. Their output is combined to make the final decision about promoter region position. TSS is then considered to be at a certain position inside the promoter region. For example, if our sliding window has length 600 bp and the positive set was extracted from -200 bp to +400 bp, then the TSS will be located at position 201 inside the predicted promoter region.

2.4 Negative set construction

When constructing the prediction model to classify promoters we need to choose what sequences to use for non-promoters. This problem is very important because it affects what features our model will use to separate the two classes. For example, suppose we choose random DNA sequences for the negative set. In this case, a very small number of them will have TATA motif at the specific position. Then the neural network model will just use this one feature to achieve almost perfect separation between the two classes. When applying such a model to real world data, the sensitivity will be high however there will be a lot of false positives. Any sequence with a TATA motif at the specific position will most likely be classified as a promoter. Simply increasing the negative set size is not an effective

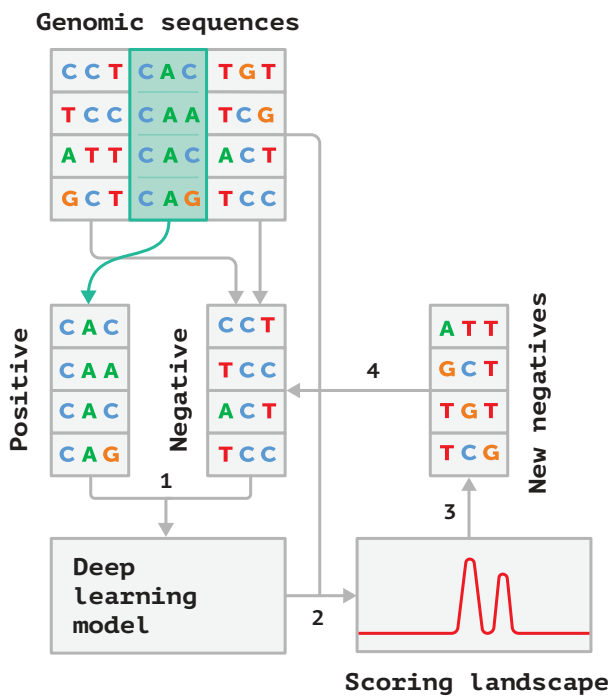


Fig. 3. Diagram of our iterative training procedure. See text for the description of each step.

solution as well, because firstly our data becomes unbalanced and secondly, there will be a big chance that neural networks will be stuck at some local minimum as in the case considered above. There are not many sequences in the negative set that will have a good scoring TATA motif, which makes our network likely to derive its recognition model heavily based on this single discriminating feature.

To resolve these issues we propose an iterative approach described below. Firstly, we choose a negative set randomly. Then we repeat the following steps:

1. We train a model with the current negative set.
2. The model is applied to the dataset with long genomic sequences and false positives are recorded.
3. A subset of false positives with the highest scores given to them by the model (the ones that are most similar to the true promoters) from each long sequence are chosen for the new negative set.
4. A new negative set is then constructed by merging the previous one with the new false positives.

This procedure is repeated until there are only a few false positives found processing the training set in step 2. These steps are illustrated in Figure 3. Such a procedure constructs a difficult negative set which helps to force our neural network to learn deeper and less obvious features to recognize a promoter sequence.

2.5 Selecting length and location of input

We need to chose what part of a promoter region to feed into our model for training. In our previous work on promoter identification we used region from -200 bp to +50 bp to extract promoter features. Since multiple transcription start points (Wang *et al.*, 2017; Dreos *et al.*, 2016) often significantly enlarge potential gene promoter regions, in this work we decided to create a promoter model using a much wider region from -1000 bp to +500 bp and then apply random substitution procedure to

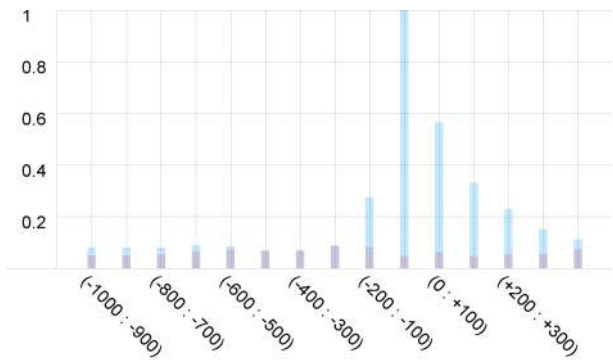


Fig. 4. Influence of different regions inside the promoter on the final score produced by the deep learning model. Blue color represents decrease of the score after random substitution and red color shows its increase.

study the location of sequence elements affecting the promoter prediction performance and potentially narrow the region down. The random substitution procedure works as follows. We have a window of size 100, which we move along each sequence with step size 100. At each position we replace the nucleotides with random 100 nucleotides and calculate new promoter score for the modified sequence. The difference between the original score and the new one is recorded and reported for each position (Figure 4). We noticed that the region from -200 bp to +400 bp has the most significant effect on the score predicted by our model and this is why it was used to train our final model.

2.6 Performance measures

In order to evaluate our method and to objectively compare predictions by our models and the other promoter identification methods, we measured performance using recall, precision, and F1 score:

$$Recall = \frac{TP}{TP+FN},$$
$$Precision = \frac{TP}{TP+FP},$$
$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision+Recall}.$$

If we predict a promoter with the TSS which is closer to the known TSS than the allowed margin for error (500 bp) then this prediction is counted as a TP. If there is no prediction in the area from -500 bp to +500 bp of the known TSS then we count this case as a FN. Any prediction outside the region from -500 bp to +500 bp of some TSS is counted as a FP. The same rule is applied for performance evaluation of all the tested promoter prediction programs. Also, we used two accuracy measures that are useful to evaluate the performance of promoter prediction tools when analyzing long genomic sequences: the average prediction error per correctly predicted TSS and the average prediction error per 1000 bp.

3 RESULTS AND DISCUSSION

3.1 Comparison of predictive performance

We compared our method to all the promoter identification tools we could obtain. A number of promoter prediction methods have been proposed. TSSW (Salamov and Solovyev, 1997) uses a linear discriminant function combining a TATA-box score, triplet preferences around the TSS, hexamer preferences and potential transcription factor binding sites. It has shown good results in a review paper by Fickett (Fickett and Hatzigeorgiou, 1997). FPROM was created by extending the TSSW program feature set, which resulted in significant improvement over TSSW and other promoter

Table 1. Comparison of the performance of the different promoter prediction methods on the test set. Results without marginal predictions (not counting original promoter predictions with low probability) are marked with asterisk.

		DeeReCT-PromID	PromCNN	PromCNN*	FPROM	FPROM*	TSSW	Promoter 2.0
Recall	TATA+	0.715	0.884	0.700	**0.908**	0.647	0.691	0.845
	TATA-	0.745	**0.948**	0.889	0.868	0.764	0.775	0.810
	BOTH	0.741	**0.940**	0.865	0.873	0.749	0.764	0.814
Precision	TATA+	**0.783**	0.118	0.242	0.236	0.491	0.252	0.107
	TATA-	**0.758**	0.127	0.320	0.227	0.476	0.259	0.104
	BOTH	**0.761**	0.126	0.310	0.228	0.478	0.258	0.105
F1 score	TATA+	**0.747**	0.208	0.360	0.375	0.558	0.369	0.190
	TATA-	**0.751**	0.224	0.471	0.360	0.587	0.388	0.184
	BOTH	**0.751**	0.222	0.456	0.362	0.584	0.386	0.186
Error per correct	TATA+	**0.277**	7.464	3.138	3.234	1.037	2.965	8.349
	TATA-	**0.320**	6.885	2.121	3.403	1.099	2.857	8.581
	BOTH	**0.314**	6.953	2.225	3.381	1.092	2.869	8.551
Error per 1000 bp	TATA+	**0.020**	0.660	0.220	0.294	0.067	0.205	0.706
	TATA-	**0.024**	0.653	0.189	0.295	0.084	0.221	0.695
	BOTH	**0.023**	0.654	0.192	0.295	0.082	0.219	0.696

recognition software (Solovyev and Shahmuradov, 2003). Promoter2.0 (Knudsen, 1999) extracted promoter elements from DNA sequences and used ANN to distinguish promoters from non-promoters based on these features. DragonGSF (Bajic and Seah, 2003) also used ANN as a part of its design and considered GC content and the concept of CpG islands for promoter recognition.

Our previous promoter recognition software, PromCNN achieved good classification performance in discriminating between short promoter and non-promoter sequences (Umarov and Solovyev, 2017). Very recently PromCNN was outperformed by (Qian *et al.*, 2018) improving accuracy by about 7%. However, as in (Umarov and Solovyev, 2017), they focused on the classification performance of short sequences, instead of promoter identification given a long genomic sequence. The latter is a much more difficult problem to tackle because of the high risk of having a large number of false positives. We could not compare our new method to theirs because they did not provide a web server or a tool that would accept long genomic sequences as inputs.

Some of the tools we came across are not available anymore. There are also tools that require extra information besides sequence data as an input. Thus here we compared our method with the following methods: PromCNN, TSSW, FPROM, and Promoter 2.0. The results are shown in Table 1. Regardless of the parameters tested, DeeReCT-PromID significantly outperforms the competitors that were examined, which showed relatively good performance in previously published papers (Bajic *et al.*, 2006; Fickett and Hatzigeorgiou, 1997). Our method has F1 score higher than the best competing tool, FPROM, by 0.153. Figure 5 shows an example of the predictions made by the different promoter prediction programs on the sequence containing the promoter of the UBE3D_1 gene. We can see that our method makes no false positive predictions while still successfully finding the true TSS.

3.2 Analyzing the learned model

It is well-known that the models trained by neural networks are difficult to interpret. We tried to overcome this limitation by visualizing the trained convolutional filters. The maximum filter length we used is 15, thus we decided to find the most important 15-mers identified by our model. We found the top 1000 most influential 15-mers and built a sequence logo for them (see Figure 6). The top three most important motifs were *CCCAGGACCATGTCT*, *GCTAGGTTGTTATGT*, *GTTCCCGGCCGGTGC*, which all contain GC rich subsequences that are well known characteristics of eukaryotic promoters (Fenouil *et al.*, 2012).

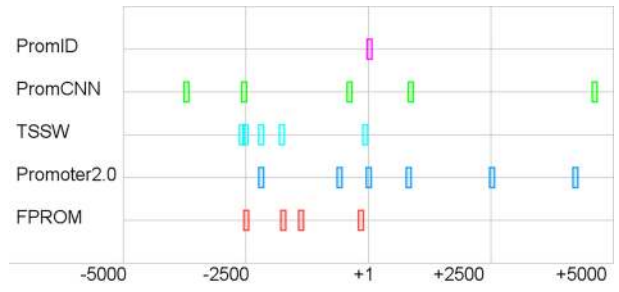
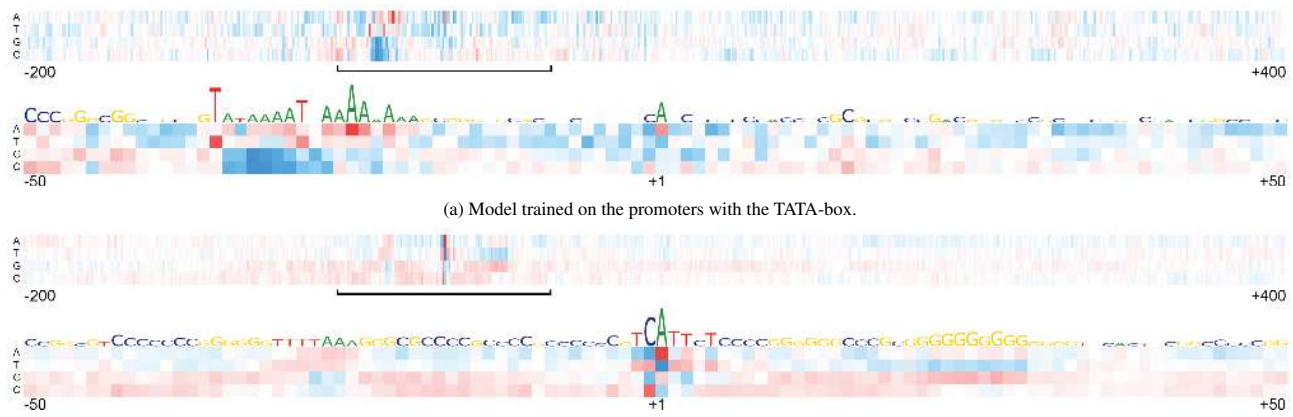


Fig. 5. Promoters predicted by the tested promoter identification programs in the DNA sequence of the UBE3D_1 gene. The true TSS is at position +1.

Fig. 6. Sequence logo of the most important 15-mers identified by our model.

To see the contributions of different nucleotides in different positions of promoter sequences we employed a modification of so called feature mutation map for all sequences in our test set. The mutation maps for TATA+ and TATA- promoters are shown in Subfigures 7a and 7b respectively. To build these maps we took a set of genomic non-promoter sequences with sizes equal to the input sequences used in our promoter models and studied how nucleotide substitutions will change the promoter score computed by our TATA+ and TATA- models. At each position of the tested sequences we replaced a nucleotide with a different one in all these sequences and computed their average promoter score. The rows in Figure 7 represent nucleotides that are used for replacement and the columns show different positions inside the promoter regions. If the new score on average increases it is represented by a red colored square. Decrease of the score is shown by using a blue colored square. The intensity of color is proportional to the effect of substitution on the score. These maps clearly illustrate significant difference of sequence features of TATA+ and TATA- promoters and location of their most conserved elements.

We can see that the largest effect on the score in TATA+ model comes from *T*/*A*-rich TATA-box region. The most significant element of TATA- promoters is the initiator element, which is very similar to the new



(a) Model trained on the promoters with the TATA-box.

(b) Model trained on the promoters without the TATA-box.

Fig. 7. Mutation map for the TSS regions from -200 bp to +400 bp. Red color represents increase of the score and blue color shows decrease.

consensus sequence for the human initiator (Inr) core promoter element *BBCABW* (where, $B = C/G/T$, $W = A/T$) described recently in (Huang *et al.*, 2017). Such initiator element typically directs the positioning of the transcription initiation start sites representing so called focused promoters in which transcription initiates at a single site or a narrow cluster of sites (Huang *et al.*, 2017). The initiator element contains a conserved motif that is observed in both (TATA+ and TATA-) data sets (see Figure 8).

For the TSS position (+1), the most preferred nucleotides are *A* and *G* . If a promoter has the initiator element then *A* is the most frequent nucleotide at position +1, otherwise it is *G*. This explains why *A* and *G* are preferred by our model at position +1 in Figure 7. The sequences at positions -1 to +3 are the most important for setting levels of basal transcription (Kugel and Goodrich, 2017). Changing nucleotides in region -30 bp to -23 bp from the original ones to *G* or *C* reduces the score considerably. While promoter regions in general have more *G* and *C* nucleotides, the mentioned region contains TATA-box in TATA+ and tends to have *T*/*A* nucleotides in TATA- promoters that is why setting nucleotides to *G* or *C* there has a negative effect on the score. In TATA- promoters we also observe occurrence of GC reach elements (Figure 7b) that is in agreement with found GC-reach most significant promoter 15-mers described above.

3.3 Accuracy of predictions

As we have shown before, promoters with the TATA-box can be predicted with a very small positional error; often the predicted TSS is exactly at the position of the true TSS. Figure 9 shows distributions of the predictions computed by our method for the test set. High positional accuracy of the TATA promoters is the result of the conserved position of several TATA promoter functional motifs relative to their TSS. However, it is not the case for promoters without the TATA-box for which the predicted positions have a normal distribution around the true TSS. For about 15% of sequences in our test set, predicted TSSs are further than 100 bp from a true TSS. This situation can be partially explained by occurrence of multiple TSSs in non-TATA promoters. Such promoters generate alternative gene isoforms that have tissue or time specific expression. It was shown in (Vo Ngoc *et al.*, 2017b) that animal promoters have focused, dispersed, and mixed transcription. In dispersed transcription, there are many weak TSSs located at the region from -50 bp to +50 bp. These multiple transcription start sites might be responsible for a wide promoter score peak (Figure 2) for non-TATA promoters generated by our deep learning model. Many of such multiple TSSs as well as some distant alternative TSSs are not annotated in the promoter databases and currently they are considered as false positives predictions while their actual status requires further experimental verification.

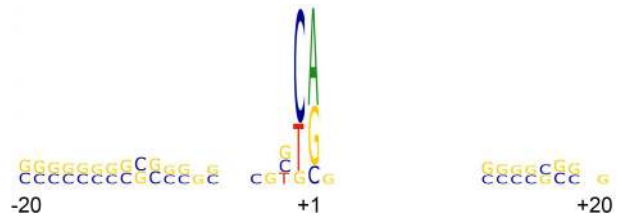


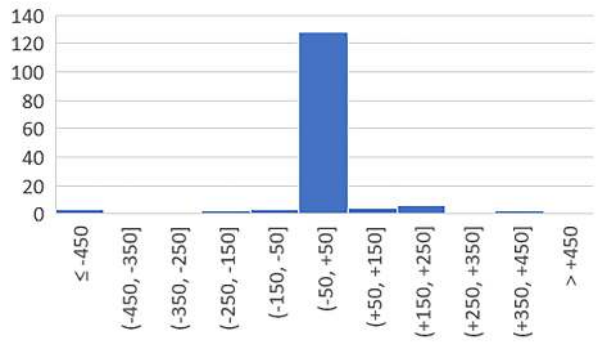
Fig. 8. Sequence logo of the region from -40 bp to +40 bp around the known TSS. The logo demonstrates sequence conservation in the promoter initiator region as well as in two GC rich regions upstream and downstream of TSS.

4 CONCLUSION

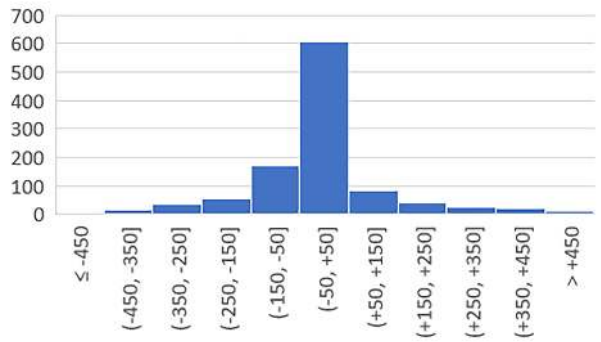
All computational promoter prediction approaches face complex organization of transcription regulation where a gene may frequently have several promoters, and within one promoter several alternative TSS locations that often are not annotated in promoter databases. All these aspects considerably complicate the development and evaluation of general promoter prediction algorithms. While previously developed promoter prediction methods can relatively accurately classify promoter and non-promoter sequences, they fail to provide good results when applied to study long genomic sequences. Due to potentially huge amount of tested locations they all have very low precision and generate a lot of false positives, which limits their usage in genome-scale studies.

In this work we have proposed a novel training technique to overcome this issue. We used iterative training that focuses on instances that were misclassified by previous iterations and builds our deep learning model that is able to eliminate the huge number of false positives. We analyzed different promoter regions to use as input for feature extraction and chose optimal input location for our tool. Evaluation of our program performance and comparing it to the available promoter prediction tools demonstrated that DeeReCT-PromID significantly outperforms other promoter finding programs.

Many genes have non-coding exons and gene-finders can not provide the actual gene start and promoter position. Therefore programs for accurate computational identification of promoters are important for revealing the gene structure and studying gene regulation. This work is



(a) Promoters with the TATA-box.



(b) Promoters without the TATA-box.

Fig. 9. Distributions of predicted TSS positions relative to the annotated TSS for the test set.

a step towards this goal while we understand that this topic is open for further investigations on structure and functioning promoter regions.

5 ACKNOWLEDGEMENTS

This work was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Awards No. FCC/1/1976-17-01, FCC/1/1976-18-01, FCC/1/1976-23-01, FCC/1/1976-25-01, FCC/1/1976-26-01, and URF/1/3412-01.

Conflict of Interest: none declared.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., *et al.* (2016). Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283.

Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T., *et al.* (2016). The ensembl gene annotation system. *Database*, **2016**.

Arnold, C. D., Zabidi, M. A., Pagani, M., Rath, M., Schernhuber, K., Kazmar, T., and Stark, A. (2017). Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. *Nature biotechnology*, **35**, 136–144.

Bajic, V. B. and Seah, S. H. (2003). Dragon gene start finder: an advanced system for finding approximate locations of the start of gene transcriptional units. *Genome research*, **13**(8), 1923–1929.

Bajic, V. B., Brent, M. R., Brown, R. H., Frankish, A., Harrow, J., Ohler, U., Solovyev, V. V., and Tan, S. L. (2006). Performance assessment of promoter predictions on encode regions in the egasp experiment. *Genome biology*, **7**(1), S3.

Butler, J. E. and Kadonaga, J. T. (2002). The rna polymerase ii core promoter: a key component in the regulation of gene expression. *Genes & development*, **16**(20), 2583–2592.

Danino, Y. M., Even, D., Ideses, D., and Juven-Gershon, T. (2015). The core promoter: At the heart of gene expression. *Biochimica et biophysica acta*, **1849**, 1116–1131.

Dreos, R., Ambrosini, G., Cavin Périer, R., and Bucher, P. (2012). Epd and epdnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic acids research*, **41**(D1), D157–D164.

Dreos, R., Ambrosini, G., Groux, R., Cavin Périer, R., and Bucher, P. (2016). The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic acids research*, **45**(D1), D51–D55.

Edwards, S. L., Beesley, J., French, J. D., and Dunning, A. M. (2013). Beyond gwass: illuminating the dark road from association to function. *The American Journal of Human Genetics*, **93**(5), 779–797.

Fenouil, R., Cauchy, P., Koch, F., Descostes, N., Cabeza, J. Z., Innocenti, C., Ferrier, P., Spicuglia, S., Gut, M., Gut, I., *et al.* (2012). Cpg islands and gc content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome research*.

Fickett, J. W. and Hatzigeorgiou, A. G. (1997). Eukaryotic promoter recognition. *Genome research*, **7**(9), 861–878.

Fredriksson, N. J., Elliott, K., Filges, S., Van den Eynden, J., StÃ¥hlberg, A., and Larsson, E. (2017). Recurrent promoter mutations in melanoma are defined by an extended context-specific mutational signature. *PLoS genetics*, **13**, e1006773.

Garieri, M., Delaneau, O., Santoni, F., Fish, R. J., Mull, D., Carninci, P., Dermitzakis, E. T., Antonarakis, S. E., and Fort, A. (2017). The effect of genetic variation on promoter usage and enhancer activity. *Nature communications*, **8**, 1358.

Huang, C. Y., Duttke, S. H., Kadonaga, J. T., *et al.* (2017). The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters. *Genes & development*.

Juven-Gershon, T., Hsu, J.-Y., Theisen, J. W., and Kadonaga, J. T. (2008). The RNA polymerase II core promoter – the gateway to transcription. *Current opinion in cell biology*, **20**, 253–259.

Kadonaga, J. T. (2012). Perspectives on the RNA polymerase II core promoter. *Wiley interdisciplinary reviews. Developmental biology*, **1**, 40–51.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Knudsen, S. (1999). Promoter2. 0: for the recognition of polii promoter sequences. *Bioinformatics (Oxford, England)*, **15**(5), 356–361.

Kugel, J. F. and Goodrich, J. A. (2017). Finding the start site: redefining the human initiator element. *Genes & development*, **31**(1), 1–2.

Lenhard, B., Sandelin, A., and Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet*, **13**(4), 233–245.

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., and Darnell, J. (2000). *Molecular cell biology. 4th*, volume 33. WH Freeman New York.

Maston, G. A., Evans, S. K., and Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*, **7**, 29–59.

Morris, J. R., Petrov, D. A., Lee, A. M., and Wu, C.-T. (2004). Enhancer choice in cis and in trans in Drosophila melanogaster: role of the promoter. *Genetics*, **167**, 1739–1747.

Qian, Y., Zhang, Y., Guo, B., Ye, S., Wu, Y., and Zhang, J. (2018). An improved promoter recognition model using convolutional neural network. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, pages 471–476. IEEE.

Rojano, E., Seoane, P., Ranea, J. A. G., and Perkins, J. R. (2018). Regulatory variants: from detection to predicting impact. *Briefings in bioinformatics*.

Roy, A. L. and Singer, D. S. (2015). Core promoters in transcription: old problem, new insights. *Trends in biochemical sciences*, **40**, 165–171.

Salamov, A. and Solovyev, V. (1997). The gene-finder computer tools for analysis of human and model organisms genome sequences. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Halkidiki, Greece*, pages 294–302.

Smale, S. T. and Kadonaga, J. T. (2003). The RNA polymerase II core promoter. *Annu Rev Biochem*, **72**, 449–479.

Solovyev, V. V. and Shahmuradov, I. A. (2003). Promh: promoters identification using orthologous genomic sequences. *Nucleic acids research*, **31**(13), 3540–3545.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, **15**(1), 1929–1958.

Taylor, M. S., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Semple, C. A. M. (2006). Heterotachy in mammalian promoter evolution. *PLoS genetics*, **2**, e30.

Umarov, R. K. and Solovyev, V. V. (2017). Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PloS one*, **12**, e0171410.

Vinagre, J., Almeida, A., PÃ³pulo, H., Batista, R., Lyra, J., Pinto, V., Coelho, R., Celestino, R., Prazeres, H., Lima, L., Melo, M., da Rocha, A. G., Preto, A., Castro, P., Castro, L., Pardal, F., Lopes, J. M., Santos, L. L., Reis, R. M.,

Cameselle-Teijeiro, J., Sobrinho-SimÃµes, M., Lima, J., MÃ¡ximo, V., and Soares, P. (2013). Frequency of TERT promoter mutations in human cancers. *Nature communications*, **4**, 2185.

Vo Ngoc, L., Cassidy, C. J., Huang, C. Y., Duttke, S. H. C., and Kadonaga, J. T. (2017a). The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters. *Genes & development*, **31**, 6–11.

Vo Ngoc, L., Wang, Y.-L., Kassavetis, G. A., and Kadonaga, J. T. (2017b). The punctilious RNA polymerase II core promoter. *Genes & development*, **31**, 1289–1301.

Wang, Y.-L., Kassavetis, G. A., Kadonaga, J. T., *et al.* (2017). The punctilious rna polymerase ii core promoter. *Genes & development*, **31**(13), 1289–1301.

Yella, V. R. and Bansal, M. (2017). DNA structural features of eukaryotic TATA-containing and TATA-less promoters. *FEBS open bio*, **7**, 324–334.

Zabidi, M. A., Arnold, C. D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., and Stark, A. (2015). Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, **518**, 556–559.