An in-depth evaluation of annotated transcription start sites in *E. coli* using deep learning

Jim Clauwaert
Department of Data Analysis and Mathematical Modelling
Ghent University
jim.clauwaert@ugent.be

Willem Waegeman
Department of Data Analysis and Mathematical Modelling
Ghent University
willem.waegeman@ugent.be

Abstract

The annotation of transcription start sites with computational methods is an important and unsolved problem in genomics. In recent years, several novel experimental methodologies – named Cappable-seq, SMRT-Cappable-seq and SEnd-seq – have been introduced for the detection of transcription start sites and applied on *E. coli*. In this study, a comparison is made between these new methodologies and the curated transcription start site data set featured by RegulonDB. The analysis between these data sets is facilitated using deep learning techniques that cover both unsupervised and supervised learning, where we expand upon a framework that allows for interpretable deep learning in genomics. This study finds annotations of recent techniques to surpass the quality of annotations provided by RegulonDB. Analysis of the transformer network trained for the detection of TSS in *E. coli* reveals its attention scores to pinpoint important promoter regions previously discussed in literature. Additionally, findings support the occurrence of a complex interaction between sense and antisense output probabilities, prevalent on key positions for interference of the transcription process.

1 Introduction

The binding region of the transcription unit, called the promoter region, is known to play an important role in the transcription rate of RNA in downstream genes. In eubacteria, the sigma factor (σ), described in 1969 [5], binds with the RNA polymerase subunit ($\alpha\beta\beta'\omega$) to create the RNA polymerase. Being part of the RNA polymerase holoenzyme, the sigma element acts as the connection between RNA polymerase and DNA and has been found to have high selectivity towards the -35 and -10 promoter elements, as positioned from the Transcription Start Site (TSS) in *E. coli* [24]. Binding of the enzyme subsequently goes with the formation of the initiation bubble (positioned from -11 to +2), placing the TSS at the active site of the polymerase [10].

The creation of annotated data sets for TSSs have ushered the development of varying methodologies [8, 36, 14]. These methods are derived from RNA-seq, a next generation sequencing methodology. A major contribution in attaining single resolution detection was realized by differential RNA-seq (dRNA-seq) [28]. dRNA-seq differentiates RNA transcripts through the $5'$ phosphate group, holding a tri-phosphate for primary transcripts and a mono-phosphate for processed transcripts, such as mature ribosomal RNA (rRNA) and transfer RNA (tRNA). Using this technique, one can create a RNA-seq library, making it possible to detect primary from processed transcripts.

Preprint. Work in progress.

More recently, several experimental procedures have been developed in quick succession for the detection and annotation of TSSs in prokaryotes: Cappable-seq, SMRT-Cappable-seq and SEnd-seq. Cappable-seq, described by Etwiller et al. [8] in 2016, is a technique aimed as an improvement on dRNA-seq through enrichment of the primary transcript. The process incorporates selective biotinylation of the $5'$ tri-phosphate ends and, unlike dRNA-seq, performs an enrichment step of the mRNA using streptavidin beads. This can lead up to a 50-fold enrichment and has indicated the existence of no less than 16,539 TSS in *E. coli*.

Single Molecule, Real Time (SMRT) sequencing is used for long read sequences and has been applied in conjunction with Cappable-seq (SMRT-Cappable-seq) [36]. This technique has allowed the sequencing of full operon constructs through capping of full length transcripts. Application of the methodology revealed the majority of annotated operons to have variations as a result of read-through at the Transcription Termination Sites (TTSs). However, one of the processing steps of the technique filters RNA reads smaller than 1,000nt.

Simultanuous $5'$ and $3'$ end sequencing (SEnd-seq), described by Ju et al. [14], is the latest developed methodology applied for TSS detection in *E. coli*. It captures both the TSS and TTS at single nucleotide resolution by circularization of the transcript using biotinylated cDNA. Fragmentation and enrichment of the biotine labeled $5'$-$3'$ fragment are followed by paired-end sequencing. The method does not introduce any selection based on RNA size. A high amount of overlapping bidirectional terminators were found within the transcriptome, suggesting a more prominent role of RNA polymerase in its regulation.

Research and development of tools to facilitate the annotation of TSSs was born using data methods for statistical inference on the nucleotide sequence of promoter sites. One of the first such published studies goes back as far as 1983, with Harr et al. [11] describing mathematical formulas on the creation of a consensus sequence. The importance of the nucleotide sequence on the prediction of TSSs – or any site of interest on the genomic sequence – is highlighted by the plethora of studies released in the last four decades. Algorithms have been designed to obtain consensus sequences [25, 17], create weight matrices [31] and rank alignments [16, 38]. Properties of interest derived from the nucleotide sequence, such as bendability [39], flexibility [3] and free energy [15] have furthermore shown importance in determining TSSs. Seq2Feature, an online tool that has been published this year, has defined as many as 41 DNA sequence-based descriptors, highlighting the ongoing interest in retrieving interpretable features from the nucleotide sequence [21].

With the advance of annotated data sets comes the use of computational methods for determining the presence of TSSs. The latest advancements in the field of machine learning and predictive genomics can be attributed to the use of deep learning methods. Convolutional neural networks were introduced for the prediction of regions with high DNA-protein affinity in 2015 [1]. In 2017, Umarov et al. further studied the use of convolutional neural networks for the annotation of TSSs [33]. More recently, we have introduced a transformer-based model for genome annotation tasks, and have achieved state-of-the-art results for the annotation of TSSs, translation initiation sites and methylation sites [6]. The introduced architecture is adapted from the transformer-XL [7], first introduced in the field of natural language processing, and is well suited to process the long nucleotide sequence data. As the transformer architecture does not imply the relative positions of the inputs w.r.t. the outputs, input data is shared between multiple output predictions and vice-versa, allowing the efficient processing of the full genome sequence. This is in contrast to the default approach where input samples are sliced from the genome, a technique that does not scale well when annotating the full genome, as the processed genome sequence length is multiplied as many times as the window size.

In this study, we have done an in-depth analysis of transcription start sites in *E. coli* by applying deep learning techniques on a set of annotated data sets obtained from varying methodologies. TSSs annotations delivered by novel techniques (Cappable-seq, SMRT-Cappable-seq and SEnd-seq) are explored alongside the annotations provided by RegulonDB. A comparative study of the data sets is performed using, among other methods, unsupervised clustering. Afterwards, we verify the use of our recently proposed transformer network on the genome annotation tasks, analyze performances between models trained on different data sets, and further explore the characteristics of model weights and its output profiles. Using both unsupervised and supervised techniques, we were able to get an interpretation on the quality of the different data sources.

The performance of the model is in line with state-of-the-art performances reported in our previous work [6]. The evaluation of the attention weights of the transformer network revealed a notable

Table 1: Overview of the data sets properties used in this study. From left to right: data sets source, *in vivo* next generation sequencing technique, amount of positive and negative labels and the partition of TSS shared with at least one other data sets. The annotations are derived or mapped (Cappable-seq) on *E. coli* MG1655 (accession: *NC_000913.3*). The total genome consists of 9,283,304 nucleotides.

data sets source	Technique	Year	Positive labels	Negative labels	Shared TSSs
RegulonDB [27]	Collection	-	6,487 (0.07%)	9,276,817 (99.93%)	2,476 (38.17%)
Etwiller et al. [8]	Cappable-seq	2016	16,348 (0.18%)	9,266,956 (99.82%)	4,631 (28.33%)
Yan et al. [36]	SMRT-Cappable-seq	2018	2,574 (0.03%)	9,280,730 (99.97%)	2,311 (89.78%)
Ju et al. [14]	SEnd-seq	2019	5,502 (0.06%)	9,277,802 (99.94%)	4,026 (73.17%)

link to the promoter regions of importance for the RNAP binding process. Based on the probability profiles of the transformer network and corroborated by previous findings of Ju et al. [14], the presence of a strong correlation between sense and antisense activity is.

2 MATERIAL AND METHODS

2.1 Data

Multiple data sets stemming from varying recent experimental methods have been considered, listed in Table 1. Different high-precision *in vivo* experimental methods were sought out and compared in order to detect or avoid structural bias.

Given its importance to the community, TSSs annotated with at least 'strong' evidence are retrieved from the RegulonDB database, summing to 6,487 annotated positions. RegulonDB features an up-to-date collection of manual and automated TSSs collected from a plethora of independent sources [27], and has been the chosen data set for recent machine learning implementations [23, 19, 35, 18].

Additional data sources involve the use of next generation sequencing techniques for TSS detection in *E. coli*. TSS annotations have been collected from studies applying Cappable-seq, SMRT-Cappable-seq and SEnd-seq as labeled data sets of the transcriptome of *E. coli*. All experiments were performed on *E. coli* and returned 16,348, 2,574 and 5,502 TSSs, respectively. Annotations of the Cappable-seq experiment, mapped using the *E. coli* MG1655 reference genome *U00096.2*, have been remapped on the *NC_000913.3* reference genome.

2.2 Model architecture

This paragraph describes the model for unsupervised clustering using autoencoders, whereas the following paragraphs in this section are on the transformer architecture for supervised classification. Clustering of the nucleotide regions surrounding the TSSs annotated by the investigated data sets has been achieved using an autoencoder combined with Principal Component Analysis (PCA) and t-SNE [20]. The autoencoder is an unsupervised learning methodology where a model is trained to optimize the encoding and decoding of the input data into and out of a compressed state, referred to as the bottleneck layer. By intuition, the higher the correlation between the input samples, the more efficiently the data can be compressed and represented by the bottleneck layer. A simple fully-connected neural network has been used, accepting a one-hot representation of a 80-nucleotide promoter sequence surrounding the annotated TSSs as input (80×4). Sequentially, the encoder part of the network is composed of a 1×1 convolutional layer with 4 kernels, transforming the input from a sparse to a dense embedding, and three fully-connected layers with 160, 96 and 8 neurons, the last of which being the bottleneck layer. The decoder of the network is trained to decode the compressed representation of the input back to its original state through three fully-connected layers of 96, 160 and 360 nodes. A one-hot representation is obtained by reshaping the output values into a 80×4 shape and applying the Softmax function over the second axis. The model is trained using the cross-entropy loss. A more detailed description of the model is given in Supplementary File 1. Figure 1a shows the architecture of the autoencoder.

For the supervised learning, a transformer model is used. The transformer model evaluates the full genome sequence by iteratively processing smaller segments s, allowing for fast training times

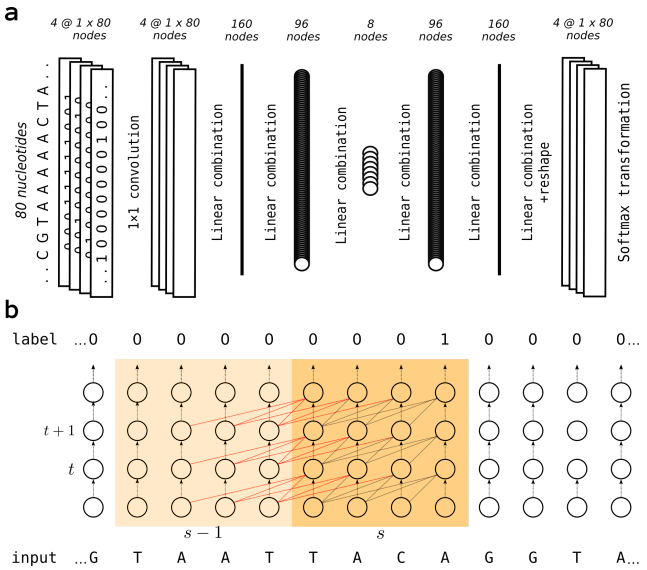


Figure 1: (**a**) Illustration of the autoencoder network used to characterize the promoter sequences of the annotated TSSs using unsupervised learning. A one-hot encoded representation of the nucleotide sequence spanning the –60 to +20 region surrounding the TSS is used. The encoder is composed of a 1×1 convolutional layer with 4 kernels and three fully connected layers with 160, 96 and 8 neurons. The decoder of the network is trained to decode the compressed representation back to the input through three fully connected layers of 96, 160 and 360 nodes. The output is reshaped into a 80×4 after which the Softmax function is applied over the second axis. (**b**) Illustration of the architecture of a simplified transformer network. The full genome is processed progressively in segments s with length l. The input sequence is sent through k consecutive layers resulting in an output probability at nucleotide resolution. $\mathbf{h}^{n,s,t+1}$, a vector representing hidden state $\mathbf{h}$ at position n in segment s of layer $t+1$, is calculated using attention to weigh the inputs of the previous layer, indicated by the connections between nodes. These hidden states are $[\mathbf{h}_1^{s,t}, ..., \mathbf{h}_n^{s,t}]$ for $n = l$ and $[\mathbf{h}^{n+1,s-1,t}, ..., \mathbf{h}^{l-1,s-1,t}, \mathbf{h}^{0,s,t-1}, ..., \mathbf{h}^{n,s,t}]$ for $n < l$. hidden states used from previous segments are fixed (red connections), making them invariant to weight updates to optimize the loss.

necessary for whole genome data processing. For every segment s with length l, the input nucleotide $x \in \{A, C, G, T\}$ is processed through k sequential layers, resulting in as many intermediary hidden states ($x \rightarrow \mathbf{h}^1 \rightarrow ... \rightarrow \mathbf{h}^k \rightarrow \hat{y}$), to obtain a prediction score $\hat{y}$ at nucleotide resolution. For each layer, hidden states are calculated using the l upstream hidden states of the previous layer. Therefore, $\mathbf{h}^{n,s,t+1}$, a vector representing hidden state $\mathbf{h}$ at position n in segment s of layer $t+1$, is computed using input vectors $[\mathbf{h}^{0,s,t}, ..., \mathbf{h}^{n,s,t}]$. The hidden states most often span multiple segments. Therefore, for $n < l$, the input vectors are more correctly represented as $[\mathbf{h}^{n+1,s-1,t}, ..., \mathbf{h}^{l-1,s-1,t}, \mathbf{h}^{0,s,t-1}, ..., \mathbf{h}^{n,s,t}]$. Hidden states made accessible from previous segments are in a fixed state that don't allow weight updates in order to minimize the loss. For each position, The output of each layer is summed with the input through a residual connection ($x \rightarrow \mathbf{h}^0 \rightarrow ... \rightarrow \mathbf{h}^{k-1} \rightarrow \hat{y}$). The propagation of information contained in hidden states (nodes) through the layers of the model is illustrated in Figure 1b, using red and black connections for hidden states from the previous or current segments, respectively.

The attention mechanism is adopted for the calculation of intermediary hidden states $\mathbf{h}^{n,s,t}$, and evaluates the relevance between two hidden states based on their vector representation. This relevance is calculated at each node and is reflected by weighing of each of the inputs hidden states of the previous layer [34]. A more in-depth explanation of the model architecture is given in Supplementary File 1 and in the article where the transformer model for genome annotations is first described [6]. Through hyperparameter tuning, a model architecture is selected (l: 512, k:6).

As the nucleotide sequence downstream of the TSS is expected to be of relevance to its identification, labels can be shifted downstream in order to make the information of these regions accessible to the

model. In this study, all labels have been shifted downstream by 20 nucleotides, a distance that is used by previous computational methods [23, 19, 35, 18]. Studies on the mechanisms of RNAP binding do furthermore not describe interactions of the RNAP binding process beyond position +20 of the TSS [13, 26]. Additionally, during hyperparameter tuning, extending the downstream sequence by shifting the labels up to 60 positions downstream did not result in better performances.

2.3 Training and evaluation

Unsupervised clustering of the data was performed using an autoencoder, PCA and t-SNE. The autoencoder has been trained using the promoter sequences of the annotated data sets (Table 1). The promoter sequence has been selected to span the region from position –60 to +20, respective to the TSS. The prediction task can be regarded as a multi-class classification problem, where the model is being optimized to correctly predict the nucleotide class (out of four) for each position of each sample. output probabilities are enforced by the softmax transformation for each of the four output classes. Therefore, the cross-entropy function is used to compute the loss and optimize the model. The loss has been weighted with respect to the data samples of each data set to ensure equal importance for each data set. All models were trained for 300 epochs using a learning rate of 0.001. Visual clustering is performed by combined dimensionality reduction on the bottleneck layer using PCA and t-SNE, sequentially reducing the dimensionality to five and two.

The transformer was trained using the full genome sequence, resulting in a total sample size of 9,283,204. The sizes of the positive and negative samples are listed in Table 1. A high imbalance exists between the positive and negative sets, the former being over four orders of magnitudes smaller than the latter. The training, test and validation set are created by splitting the genome at three (randomized) positions that constitute 70%, 20% and 10% of the genome, respectively. Each set has been evaluated to ensure the relative contributions of the positive labels to be comparable. The same splits are performed to train and evaluate all models, at positions 2,738,785, 3,667,115,950 and 4,131,280, as indexed by the *RefSeq* genome (accession: *NC_000913.3*). Both the sense and anti-sense components of these strands are thereby separated into the same sets to warrant no unforeseen information transfer – that might result in a bias – between the different sets. The minimum loss on the validation set during training is used to determine the point at which the network training is halted. Given the prediction task to be a binary classification problem, the cross-entropy loss is used. The performance metrics of these models are obtained by evaluation on the test set.

The Area Under the Receiver Operating Curve (ROC AUC) and Area Under the Precision Recall Curve (PR AUC) have been used as evaluation metrics. ROC AUC performances represent the area under the curve created by connecting the true positive (y-axis) and false positive (x-axis) rate of the predictions at different thresholds of the output probability. However, due to the imbalance between positive and negative labels, even a small false positive rate results in a set of false positive predictions that heavily outweighs the true positives in absolute count. PR AUC represents the contribution of true positives in proportion to the positives predictions (precision, y-axis) with respect to the recall of the positive set (x-axis) for different thresholds on the output probability. The use of this metric is of importance, as it gives a better depiction of the ability of the model in a practical setting, where the top-k positive predictions are of interest.

3 Results

3.1 Data exploration indicate new techniques to show improved quality annotations

Before training the transformer network to detect TSSs, the data sets were explored in order to understand the similarities and dissimilarities between the different experimental methods used to annotate the TSSs on the genome. First a comparison is made between the shared TSSs. In line with the criteria applied by [36], a TSSs is shared between two data sets if they are annotated within a distance of five nucleotides. The distribution of shared TSSs between the data sets is given in Figure 2a. The total amount of TSSs that are shared by all four data sets is 1,012. The sum of annotated TSSs shared by three or more data sets is 2,224 and by two or more 5,104. The fraction of TSSs that are present in other data sets are lowest for RegulonDB (38.17%) and Cappable-seq (28.33%) and highest for SMRT-Cappable-seq (89.78%) and SEnd-seq (73.17%). However, percentages should be considered in relation to the relative sizes of the data sets, e.g. Cappable-seq has the largest

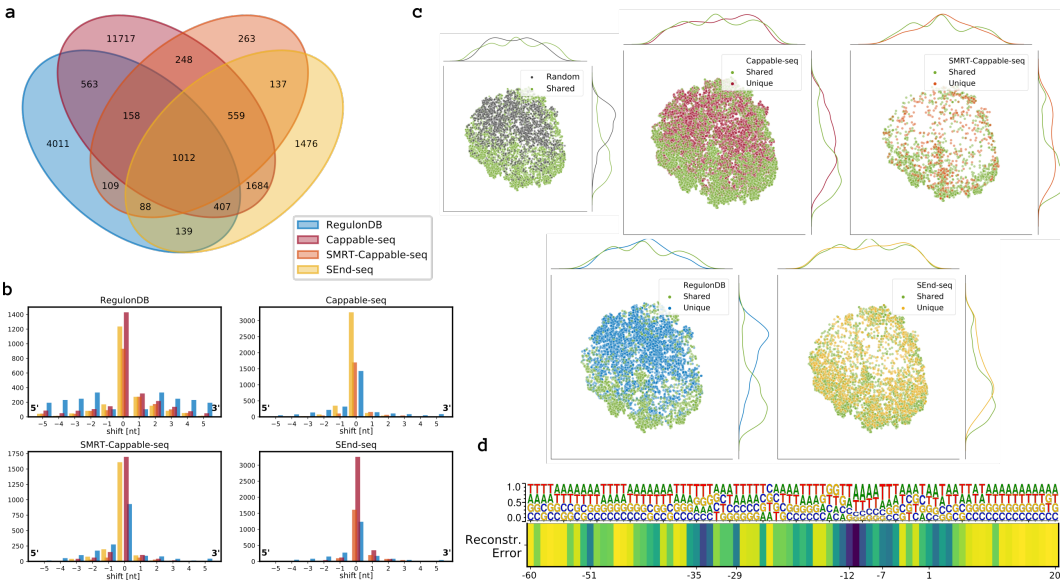


Figure 2: Comparisons between transcription start site (TSS) annotations provided by varying methodologies in *E. coli*. The annotations are obtained from RegulonDB and recent publications introducing novel methodologies from Etwiller et al. (Cappable-seq), Yan et al. [36] (SMRT-Cappable-seq) and Ju et al. [14] (SEnd-seq). (**a**) A venn diagram represents the overlap of annotated TSSs between the different data sources. TSSs are considered to be shared when annotated within 5 nucleotides of one another. (**b**) The distribution of the annotation shift existent between the shared TSSs within the 5-nucleotide distance used to denote a shared TSS. Distributions are centered on the annotated position for each data set separately. (**c**) Scatter plots of the bottleneck representation of the promoter sequence of each set of annotations after applying PCA and t-SNE. The latent distributions for each plot are given. Representations are divided into promoter sequences present in other data sets (shared) and those that are unique for that data set. Additionally, a random set of sequences has been sampled from the genome and shown alongside the representations of the promoter sequences of all shared TSSs. (**d**) The mean reconstruction error of the input sequences using increasing sizes of bottleneck layer of the autoencoder. The colormap represents lower values as darker blue and higher values as yellow. Additionally, the probability scores for each nucleotide at each position is given based on the positional frequences of the input samples on which the model is trained.

absolute amount of shared TSSs. The fraction of TSSs in each data set that has been annotated by other methods is given in Table 1.

Figure 2b shows the distributions of the distances between the TSSs that are shared within the 5-nucleotide distance. The fraction of shared TSSs annotated at single-nucleotide resolution in comparison with all those present within the 11-nucleotide window constitutes 52.9%, 51.1% and 57.0% for RegulonDB on Cappable-seq, SMRT-Cappable-seq and SEnd-seq, 52.9%, 78.7% and 83.4% for Cappable-seq on RegulonDB, SMRT-Cappable-seq and SEnd-seq, 51.1%, 78.7% and 77.7% for SMRT-Cappable-seq on RegulonDB, Cappable-seq and SEnd-seq, and 57.0%, 83.4% and 77.7% for SEnd-seq for RegulonDB, Cappable-seq and SMRT-Cappable-seq. The annotations of RegulonDB show a clear distinction with the other three data sets, ranging from 51.1 to 57.0 percent. In contrast, the precision of TSSs between Cappable-seq, SMRT-Cappable-seq and SEnd-seq is higher (77.7-83.4).

Using unsupervised clustering, it is possible to explore causes of variability between data samples. For example, samples that are clustered within their data sets are a strong indication of the applied methodology introducing a bias in the promoter sequence. After training an autoencoder using all the positive annotations of each data set, the compressed representation (i.e. bottleneck layer) of the samples is used as input for PCA and t-SNE. Figure 2c shows the resulting scatter plots and respective distributions of the samples along the two axes for each data set. TSSs that are

unique are differentiated from those that have been annotated by one of the other data sets. The incidence of a position being identified as a TSSs by two independent studies (i.e. shared TSS) serves as a validation as the chance of this to happen is exceedingly small, given the amount of positive annotations with respect to the size of the genome. Furthermore, for reference, a set of 2,500 sequences is randomly sampled from the genome and is plotted alongside all shared TSSs of all data sets.

No clear formation of separate clusters is formed. Still, clear differences do exist between the space occupied by TSSs that are unique to the data set and those that are shared. The space occupied by the shared TSSs in all datasets follows the same distributions. RegulonDB shows the least overlap between the unique and shared TSSs, where the annotations unique to the set are most consistent with the distributions of randomly sampled sequences. In contrast, TSSs uniquely present in SEnd-seq shows distributions that are most similar to those that have been validated by other methodologies, a trend that is correlated to the novelty –by year– of each technique.

Lastly, multiple autoencoders have been trained on the data using varying sizes ofs the bottleneck layer. By evaluation of the loss on the reconstructed sequence (i.e. reconstruction loss), it is possible to differentiate between the error for each of the positions along the reconstructed sequence. Based on the properties of the autoencoder, the model will minimize the reconstruction loss for promoter regions indicative of the presence of a TSS, as these nucleotide positions are expected to show high correlations between the positive samples. This is in contrast to nucleotide regions that have no influence on the occurence of a TSSs, for which compression is not possible. Through evaluation of the reconstruction loss for autoencoders with varying sizes of the bottleneck layer, which restricts the amount of information available to compact the sequence, nucleotide positions indicative of the presence of a TSS can be ranked by importance. Figure 2d gives the mean reconstruction error off all the annotated TSS for different sizes of the bottleneck layer (4,8,12,16,24,32), alongside the positional frequencies of the input samples on which the model was trained. In contrast to a position weight matrix, returning no discernable motif, the autoencoder is able to map higher-order interactions between the nucleotides. The individual reconstruction losses at each position for all of the models is given in Supplementary Figure S1.

A curated annotation using all four data sources is made in order to combine the provided information and to compare it in a supervised learning setting. The data exploration gave insights into the data properties, and has resulted in the following criteria for the construction of the custom data set: (1) the TSS is detected by at least two separate data sets, (2) the exact position of two TSSs within a 5-bp window is determined by the position supported by the majority of data sets and in case of a tie, (3) the position is selected based on the novelty of the technique (SEseq $>$ SMRT-Cappable-Seq $>$ cappable-seq $>$ RegulonDB). Finally, given the novelty of SEnd-seq and the similarity of the clusters of the samples that are both shared and unique to this set (shown in Figure 2c), all TSSs from this data set are added to the custom set. The resulting custom data set has a total of 6,580 positively labeled positions. The described criteria do not guarantee the optimal outline for the combination of the different sets of annotated TSSs, nor does it give any guarantees about the quality of the different methodologies. In fact, lacking annotations that ascertain the ground truth, it is unclear how different approaches would be validated. Instead, the methods applied here, such as unsupervised clustering, can only offer an indication on the state and relatedness of the discussed data sets.

3.2 Improved model performances are obtained by combination off all available annotation sets

As discussed, the validation and test set used for each model span identical regions of the genome. Therefore, trained models can easily be evaluated using the annotations of investigated methods. Each set of annotations has been used for training and evaluation purposes. Table 2 gives a full overview of the performances of all the models on each of the test sets. The transformer networks trained on the custom TSSs and the set derived from SEnd-seq show the highest predictive performances on their respective test sets with a ROC AUC of 0.976 and 0.978. The models trained on SMRT-Cappable-seq and Cappable-seq annotations both have similar performances of 0.958 and 0.960. Data from RegulonDB displays the least predictability, having a ROC AUC no higher than 0.882. Interestingly, the transformer trained on the custom annotations performs best on the custom test set, and those annotated by both SMRT-Cappable-seq and SEnd-seq.

Table 2: Performance metrics of the transformer network trained and tested on the labeled transcription start sites obtained from RegulonDB [27], Etwiller et al. [8] (Cap(pable)-seq), Yan et al. [36] (SMRT-(Cappable-)seq), Ju et al. [14] (SEnd-seq) and the curated custom set. A model has been trained (rows) and evaluated (columns) on each set of annotations. Both Area Under the Receiver Operating Curve (ROC AUC) and Area Under the Precision Recall Curve (PR AUC) are given for each set-up.

Train set			**Test set**		
	RegulonDB	Cap-seq	SMRT-seq	SEnd-seq	Custom
ROC AUC					
RegulonDB [27]	**0.882**	0.815	0.923	0.882	0.885
Cap-seq [8]	0.790	**0.961**	0.938	0.945	0.945
SMRT-seq [36]	0.749	0.899	0.958	0.961	0.956
SEnd-seq [14]	0.669	0.835	0.944	0.978	0.964
Custom	0.740	0.920	**0.976**	**0.981**	**0.976**
PR AUC					
RegulonDB [27]	0.030	0.026	0.053	0.057	0.064
Cap-seq [8]	0.014	**0.132**	0.029	0.039	0.044
SMRT-seq [36]	0.029	0.044	0.086	0.081	0.089
SEnd-seq [14]	0.035	0.052	0.098	**0.128**	0.137
Custom	**0.039**	0.057	**0.141**	**0.128**	**0.141**

Evaluating the results in function of the test sets, RegulonDB shows the strongest difference between the best and worst performing model. The custom set, SMRT-Cappable-seq and SEnd-seq show the least variability. Although the performances of the different models on these test sets show high similarity with one another, the trained model of the custom training set distinguishes itself. Only 2,574 annotated TSSs were obtained through SMRT-Cappable-seq, an amount that is lower than the expected amount of TSSs present on the genome of *E. coli*. Nonetheless, the data set provided by SMRT-Cappable-seq shows the highest percentage of shared TSSs. The substantial improvement of ROC AUC achieved by the custom model on the test set of SMRT-Cappable-seq can be explained by a lower occurrence of false negative predictions in the models predictions. Reducing the amount of false negative predictions improves the ROC AUC score even if at the cost of false positive predictions, as the true positive rate (y-axis) is more sensitive to change than the false positive rate (x-axis) due to the different sizes of the denominator (i.e. class imbalance).

The low scores of the PR AUC metric illustrate that, in contrast to the high performances for ROC AUC, the absolute amount of false positive labels still outweighs the true positive predictions, resulting in a low precision/recall ratio of the positive annotations. The ROC/PR AUC curves of the model trained on the custom annotation on each of the test sets is given in Supplementary Figure S2 and S3.

3.3 The attention layers of the transformer network pinpoint key promoter regions

The attention mechanism is used by the model to selectively collect information for the calculation of the hidden states of the next layer. Through attention, a weight is assigned to the upstream hidden states $[\mathbf{h}^{n-l,t}, ..., \mathbf{h}^{n,t}]$ for the calculation of $\mathbf{h}^{n,t+1}$. In natural language processing, the evaluation of attention weights in trained models has shown the connection of semantically-relevant words [34]. Figure 3 displays the weights of the the six attention heads averaged over the six different layers of the model at the positively labeled positions of the custom test set. As the segment length of the model is 512 nucleotides, as many weights are given for each position ranging from positions –492 to +20 w.r.t. the annotated TSS – as labels have been shifted 20 nucleotides downstream before training. The featured heat maps display only the last 100 weights of the model. The full mapping of the 512 value weight vector, in line with the segment length used during training, is given by Supplementary Figure S4.

Attention shows to successfully pinpoint sites of relevance within the promoter region. Overall, we see high relevance of the around the TSSs, +21 site and –35 and –10 box, two hexamer regions spanning the promoter sequence from –35 to –29 and –12 to –7 (attention head 1,2,4,6). Attention head 1 and 2 focus on nucleotide bases ranging from –11 and –7, which interact with region 2 of σ^{70} (σ_2) during isomerization of RNAP as it forms the open complex [26]. Nucleotides at the –6/–5 sites have affinity to $\sigma_{1.2}$ [9, 12], focused on by attention head 2. Barne et al. describes the interaction of region $\sigma_{2.5}$ of the RNAP enzyme with the extended –10 motif –nucleotides -14/-13– [4], a region

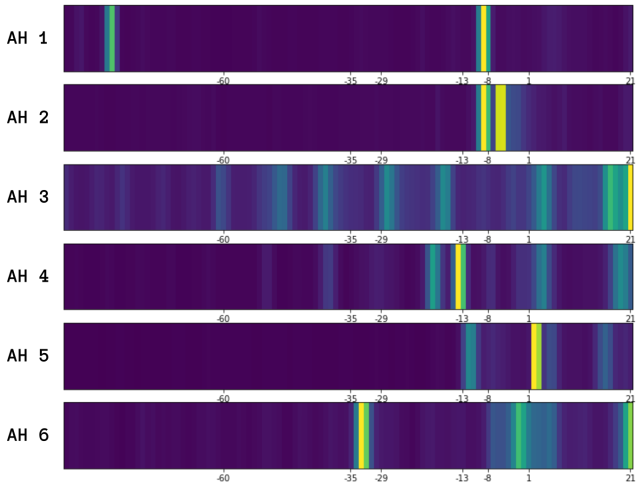


Figure 3: Attention of the model trained and evaluated on the custom annotations for the six attention heads (AH). The attention weights are given for all positive samples in the test set and are averaged over the six layers. Displayed attention scores span from a region ranging from –100 to +21 with respect to the transcription start site. Attention is in correlation with the contribution of each position to the output of the attention head. The colormap represents low values as darker blue and higher values as yellow. Given results substantiate the ability of the model to mark regions of importance towards the binding of the RNA polymerase and, therefore, transcription start site.

highlighted by attention head 4. Attention head 5 features the boundaries of the open complex, recently described by Henderson et al. to extend over the region of delimited by positions –11 to +2 for λP_K promoters and –12 to +2 for T7A1 promoters [13]. Lastly, the –35 box, highlighted by attention head 6, interacts with σ_4 [30].

The large weight assigned to position +21 site is expected, given the residual connection between between $\mathbf{h}^{n,t}$ and $\mathbf{h}^{n,t+1}$. Other sites, such as position –80 by attention head 1 or those given by attention head 3, do not offer a straightforward interpretation or connection to the literature.

3.4 The output probabilities of the model reflect a complex interaction of sense and antisense

Using the model to predict the occurrence of a TSSs on the test set results in a continuous output probability profile along the genome. In this section, the transformer network trained on the custom annotation has been evaluated alongside the annotations of the coding sequences, provided by the *RefSeq* assembly. The median value of the output probability over the full test set equals 0.0052, and serves as a baseline. The median metric is used in place of the mean as it is less sensitive to outliers (see supplementary Figure S5).

The output profile of the model reveals broad areas of increased probabilities for the sense strand around TSSs. Surprisingly, this is often in correlation with increased activities on the antisense strand. Figure 4a shows the median values at each postion around the TSSs of the test set. The outline shows the increased output probability on both the sense and antisense strand, where the latter does not only show overall increased activity but also peaks upstream of the TSSs.

Analysis of the probability profile reveals an overall increased output probability for positions in between regions for which no coding sequence is present on the forward and reverse strand. The median probability over all regions labeled as a coding sequence is 0.0049, and is in line with the median over the full test set. For regions in between coding sequences on either strand, the median is doubled from the baseline (0.0105).

A high probability is not only present for bordering outward-facing coding sequences. sense and antisense activity is prevalent for bordering coding sequences within operons. Using the operon mapping of RegulonDB, the median probability within the non-coding sequences of operons is

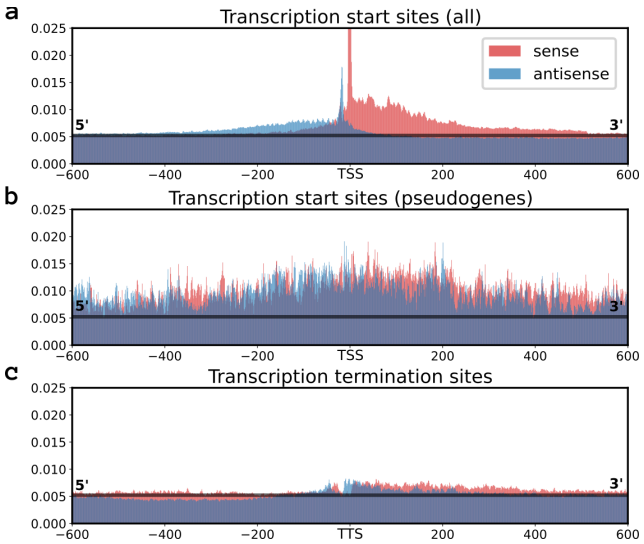


Figure 4: An overview of several plots that describe the prediction characteristics of the model. Data is obtained by the model trained and evaluated on the custom test set. the median output value over the full test set of 0.053 is represented as a horizontal black line in (a,b,c). Each plot gives the median model output for all positions within a 601-nucleotide window centered on the site of interest. Displayed are the median model outputs for the positions around all the (**a**) Transcription Start Sites, (**b**) the TSSs of genes annotated as pseudogenes and (**c**) all the transcription termination sites (TTS), obtained from Ju et al. [14].

0.0079 for sense and 0.0071 for antisense. These short sequences have a large amount of peaks, where the median over the maxima of each region is ca. ten times the baseline with 0.0415 and 0.0480 for sense and antisense, respectively.

Transcription termination sites were obtained from Ju et al. [14] and shown in Figure 4c. The median scores figure shows a slightly increased probability score around the TTS for both strands with a shifted dip at the center.

Lastly, pseudogenes were found to encompass areas with a substantial increase of output probability along both coding and non-coding nucleotide sequences, resulting in an increase of false positives. The median output probability within the coding sequences of pseudogenes is 0.0111. Most interestingly, the profile around these regions is unique, as no patterns can be discerned. Figure 4b gives the median values for each position around the TSSs of coding sequences annotated as pseudo-genes. Figure 4c gives an example of a pseudogene region.

Supplementary Figure S6a-c show the model output along the custom TSSs annotations as displayed on the UCSC browser (see Supplementary Information). The figure shows examples of the close interaction between the sense and antisense of the model output. The transcription profile from the SMRT-Cappable-seq experiment is also displayed.

4 Discussion

The creation of a high-performing predictive model for the annotation of TSSs is of high interest to the scientific community where it can, for example, aid in a better understanding of various biological processes. However, no substantial progress in the detection of TSSs in prokaryotes has been forthcoming in recent years. The prediction tasks at hand suffer from several well-known hurdles in the application of machine learning, such as (i) imbalanced classes, presenting a negative set that is three orders of magnitude larger than the positive set, (ii) the lack of annotation and (iii) a high amount of noise inherent to applied methods. More specific to this setting are (iv) the obstacles introduced by the dimensions of the input feature space, a multi-million length sequence, and (v) the biological variation of the transcription landscape, which is dependent on, among other factors, the growth phase and conditions of the organism.

Activation of transcription factors is commonly dependent of growth conditions and results in the expression of different sets of genes. Given the fact that all discussed methodologies are based on the use of expression data, variability is to be expected. SEnd-seq combines results from both the exponential and log growth phase, SMRT-Cappabe-seq has been executed during the late-log growth phase and Cappable-seq during the mid-log growth phase. Consequently, the detection of TSSs activated by specialized σ-factors, such as σ^{38} for stress responses, σ^{32} for heat shock response and σ^{28} associated with flagellar genes [32], is bound to be limited. Due to limitations to the existing methods, it is not possible to define which of the seven well-defined sigma factors in *E. coli* are associated with each TSS. It can be expected that the used data sets mainly hold TSSs activated by the household σ^{70}-factor. The creation of high-quality annotations for each of the σ-factors is required to create a better understanding on the relation between σ-factors and their promoter sites.

Differences between the annotations obtained from recently developed methodologies highlight the fact that single-nucleotide resolution is not guaranteed. The higher amount of shifts in the annotations from RegulonDB, where the majority (80%) of the positions are described to be obtained through dRNA-seq, underline this technique – and data set – to be of lower quality. This is confirmed by unsupervised learning, where the promoter sequences from TSSs uniquely present in RegulonDB show the highest conformity with random sequences from the genome (see Figure 2b). Here, promoter sequences from SMRT-Cappable-seq and SEnd-seq, the two most recent techniques, were shown to be of the highest quality.

When comparing the performance metrics of the transformer network (see Table 2, we can see that irregularities between these datasets are propagated, reflected by their respective performances on the individual test sets. A custom positive set is created derived from the available data sets that is used to train the network. Overall, this transformer model returns the best performances on the variety of test sets, and has been used for further analysis.

The results of both the supervised and unsupervised learning display a considerable discrepancy between the annotations of different methodologies. Only the annotations from SMRT-Cappable-seq and SEnd-seq, even though methodologically very different, show comparable characteristics. Until today, investigation of TSSs or promoter sequences is performed using the annotations from RegulonDB, a choice that might not be justified, given the results of our study. As the majority of annotations of RegulonDB is based on dRNA-seq, a revision of the database might be necessary, based on findings of more recent methodologies. In this study, the use of all data sets has proven beneficial towards the predictive performance of the model.

Deep learning methodologies have become a popular choice in learning from nucleotide sequences thanks to several advantages, such as the automatic extraction of features and the capacity to handle big data. Especially the use of convolutional layers have resulted in several successful applications achieving state-of-the-art results in recent years [1, 37, 2]. However, conventional machine learning methodologies are not used to process the full genome. Instead, the genome around the feature of interest is sliced to create separate input samples. Processing the full genome results in a multiplication of the input data that is dependent on the size of the window size, and is therefore rarely done.

We recently introduced a custom transformer architecture for processing the full genomic sequence and obtained state-of-the-art performances for the annotation of TSSs, translation initiation sites and methylation sites [6]. Importantly, the transformer architecture does not assert the relative positions of the input nucleotides with respect to the output label. Inputs are only processed once, and intermediate values, calculated from multiple inputs, are shared between multiple outputs. Therefore, the cost of extending the window size, i.e. increasing the segment length l, is reduced. In comparison to recurrent neural networks, the transformer architecture is not hindered by slow training times as an effect of horizontal data transfer.

Using this novel transformer architecture on data from discussed methodologies for TSSs detection, ROC AUC scores as high as 0.981 were obtained. Due to the existence of a large amount of false positives, the resulting PR AUC is no higher than 0.14. The low scores of PR AUC emphasize the use of the model as-is to be impractical. However, the output of the model displays multi-nucleotide peaks, an effect most probably resulting from the uncertainty introduced by the different data sets that show variations on single-nucleotide level (Figure 2). Reducing the resolution of the prediction task towards 4-mers drastically improves the PR AUC to 0.29. The PR AUC scores stabilize for further reduction of the resolution, e.g. 10-mers result in a PR AUC of 0.33.

Analysis of the attention weights has revealed conformity between regions important for the detection of TSSs and those described in literature. This remarkable connection serves as an important advantage that has not been observed in other deep learning methodologies for genome annotation tasks. The correlation of the attention weights of the network to regions of importance can speak for both the validation of the model and the guidance of future studies.

Analysis of the model output reveals increased probabilities to be more common at specific locations along the genome, where interaction along the sense and antisense strand corroborate mechanisms of interference as described in the literature. Several types of transcription interference exist that involve convergent promoters: promoter competition, sitting duck, occlusion, collision and roadblock interference [29]. Both sitting duck and roadblock interference involve the occurrence of a RNAP polymerase open complex that is too weak or strong for promoter escape. Promoter competition involves the inhibition of bordering promoters while occlusion and collision are related to the inference of head-on elongating RNAP. Recently, through the analysis of the SEnd-seq data, the occurrence of non-coding RNA is described to play an under-appreciated role in head-on transcription interference [14].

The abundance of high probability scores along the genome sequence in relation with TSSs or other key positions for promoter interference or both give a likely explanation for the high amount of false positive predictions. RNAP binding sites or promoter sequences that can not be detected by *in vivo* methodologies are bound to be present on the genome. These include sequences with an affinity too high for promoter escape, those inhibited by interference/inhibition mechanisms and TSSs that result in RNAs too small for detection. For example, the low transcriptional activity common for pseudogenes [22] could be explained by the interfering affinity of the RNA polymerase for both sense at antisense, as indicated by the probability profile of the model output (see Figure 4b and Supplementary Figure 6c).

While the model might be performant in the detection of promoter sites, it lacks the information required to differentiate between those promoter sites that result in detectable transcribed RNA (i.e. a TSS) and those that do not. This presents an extra challenge for predictive tools that is a likely explanation towards the lower predictive performance of TSSs in comparison with translation initiation sites, where a ROC AUC of 0.998 is obtained [6]. With respect to RNA polymerase interference, post-processing of the model output profiles for both strands could be viable for further improvement of the predictive model.

The advent of improved methodologies for the detection of genomic loci and the development of novel machine learning techniques offer new opportunities for predictive models towards genome annotations tasks. TSSs show to be an example where a new level of complexity is revealed from the interplay between sense and antisense by expanding the learning problem to include the full genome. The generated genome-wide results from the discussed machine learning algorithms highlight the capability of these techniques to offer a better understanding into the workings of the predictive model, the data it is trained on and the biological process related to the prediction task.

5 SUPPLEMENTARY DATA

All models and code used is available on GitHub. `https://github.com/jdcla/DNA-transformer`. Supplementary File 1 features the full technical details of the transformer network and Supplementary Figures. The annotations and model outputs are given in Supplementary File 2. All annotations and output of the model trained on the custom set can be browsed on UCSC (load session: `https://kermit.ugent.be/files/UCSC/UCSC_browser.html`).

6 ACKNOWLEDGEMENTS

The authors acknowledge the support of Ghent University.

7 FUNDING

J.C. is supported by the Special Research Fund (BOF24j2016001002) from Ghent University. This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

7.0.1 Conflict of interest statement.

None declared.

References

[1] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. 33(8):831–838. ISBN: 1087-0156 1546-1696.

[2] C. Angermueller, H. J. Lee, W. Reik, and O. Stegle. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biology*, 18(1):67, Apr. 2017.

[3] D. G. Ayers, D. T. Auble, and P. L. deHaseth. Promoter recognition by escherichia coli RNA polymerase: Role of the spacer DNA in functional complex formation. 207(4):749–756.

[4] K. A. Barne, J. A. Bown, S. J. Busby, and S. D. Minchin. Region 2.5 of the escherichia coli RNA polymerase sigma70 subunit is responsible for the recognition of the 'extended -10' motif at promoters. 16(13):4034–4040.

[5] R. R. Burgess, A. A. Travers, J. J. Dunn, and E. K. F. Bautz. Factor stimulating transcription by RNA polymerase. 221(5175):43–46.

[6] J. Clauwaert and W. Waegeman. Novel transformer networks for improved sequence labeling in genomics. *bioRxiv*, page 836163, Nov. 2019.

[7] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context.

[8] L. Ettwiller, J. Buswell, E. Yigit, and I. Schildkraut. A novel enrichment strategy reveals unprecedented number of novel transcription start sites at single base resolution in a model prokaryote and the gut microbiome. 17(1):199.

[9] A. Feklistov, N. Barinova, A. Sevostyanova, E. Heyduk, I. Bass, I. Vvedenskaya, K. Kuznedelov, E. Merkienė, E. Stavrovskaya, S. Klimašauskas, V. Nikiforov, T. Heyduk, K. Severinov, and A. Kulbachinskiy. A basal promoter element recognized by free RNA polymerase sigma subunit determines promoter recognition by RNA polymerase holoenzyme. 23(1):97–107.

[10] T. J. Gries, W. S. Kontur, M. W. Capp, R. M. Saecker, and M. T. Record. One-step DNA melting in the RNA polymerase cleft opens the initiation bubble to form an unstable open complex. 107(23):10418–10423.

[11] R. Harr, M. Häggström, and P. Gustafsson. Search algorithm for pattern match analysis of nuleic add sequences. 11(9):2943–2957.

[12] S. P. Haugen, M. B. Berkmen, W. Ross, T. Gaal, C. Ward, and R. L. Gourse. rRNA promoter regulation by nonoptimal binding of sigma region 1.2: An additional recognition element for RNA polymerase. 125(6):1069–1082.

[13] K. L. Henderson, L. C. Felth, C. M. Molzahn, I. Shkel, S. Wang, M. Chhabra, E. F. Ruff, L. Bieter, J. E. Kraft, and M. T. Record. Mechanism of transcription initiation and promoter escape by e. coli RNA polymerase. 114(15):E3032–E3040.

[14] X. Ju, D. Li, and S. Liu. Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria. 4(11):1907–1918.

[15] A. Kanhere and M. Bansal. A novel method for prokaryotic promoter prediction based on DNA stability. 6(1):1.

[16] C. E. Lawrence and A. A. Reilly. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. 7(1):41–51.

[17] C. Lefèvre and J.-E. Ikeda. Pattern recognition in DNA sequences and its application to consensus foot-printing. 9(3):349–354.

[18] H. Lin, Z. Liang, H. Tang, and W. Chen. Identifying sigma70 promoters with novel pseudo nucleotide composition. pages 1–1.

[19] B. Liu, F. Yang, D.-S. Huang, and K.-C. Chou. iPromoter-2l: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. 34(1):33–40.

[20] L. v. d. Maaten and G. Hinton. Visualizing data using t-SNE. 9:2579–2605.

[21] R. Nikam and M. M. Gromiha. Seq2feature: a comprehensive web-based feature extraction tool. 35(22):4797–4799.

[22] T. T. Perkins, R. A. Kingsley, M. C. Fookes, P. P. Gardner, K. D. James, L. Yu, S. A. Assefa, M. He, N. J. Croucher, D. J. Pickard, D. J. Maskell, J. Parkhill, J. Choudhary, N. R. Thomson, and G. Dougan. A strand-specific RNA–seq analysis of the transcriptome of the typhoid bacillus salmonella typhi. 5(7).

[23] M. S. Rahman, U. Aktar, M. R. Jani, and S. Shatabda. iPro70-FMWin: identifying sigma70 promoters using multiple windowing and minimal features. 294(1):69–84.

[24] M. Rosenberg and D. Court. Regulatory sequences involved in the promotion and termination of rna transcription. 13(1):319–353.

[25] M. A. Roytberg. A search for common patterns in many sequences. 8(1):57–64.

[26] R. M. Saecker, M. T. Record, and P. L. deHaseth. Mechanism of bacterial transcription initiation: RNA polymerase - promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis. 412(5):754–771.

[27] A. Santos-Zavaleta, H. Salgado, S. Gama-Castro, M. Sánchez-Pérez, L. Gómez-Romero, D. Ledezma-Tejeida, J. S. García-Sotelo, K. Alquicira-Hernández, L. J. Muñiz-Rascado, P. Peña-Loredo, C. Ishida-Gutiérrez, D. A. Velázquez-Ramírez, V. Del Moral-Chávez, C. Bonavides-Martínez, C.-F. Méndez-Cruz, J. Galagan, and J. Collado-Vides. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in e. coli k-12. 47:D212–D220.

[28] C. M. Sharma, S. Hoffmann, F. Darfeuille, J. Reignier, S. Findeiß, A. Sittka, S. Chabas, K. Reiche, J. Hackermüller, R. Reinhardt, P. F. Stadler, and J. Vogel. The primary transcriptome of the major human pathogen helicobacter pylori. 464(7286):250–255.

[29] K. E. Shearwin, B. P. Callen, and J. B. Egan. Transcriptional interference – a crash course. 21(6):339–345.

[30] U. Siebenlist, R. B. Simpson, and W. Gilbert. E. coli RNA polymerase interacts homologously with two different promoters. 20(2):269–281.

[31] G. D. Stormo and G. W. Hartzell. Identifying protein-binding sites from unaligned DNA fragments. 86(4):1183–1187.

[32] L. G. Treviño-Quintanilla, J. A. Freyre-González, and I. Martínez-Flores. Anti-sigma factors in e. coli: Common regulatory mechanisms controlling sigma factors availability. 14(6):378–387.

[33] R. K. Umarov and V. V. Solovyev. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. 12(2):e0171410.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need.

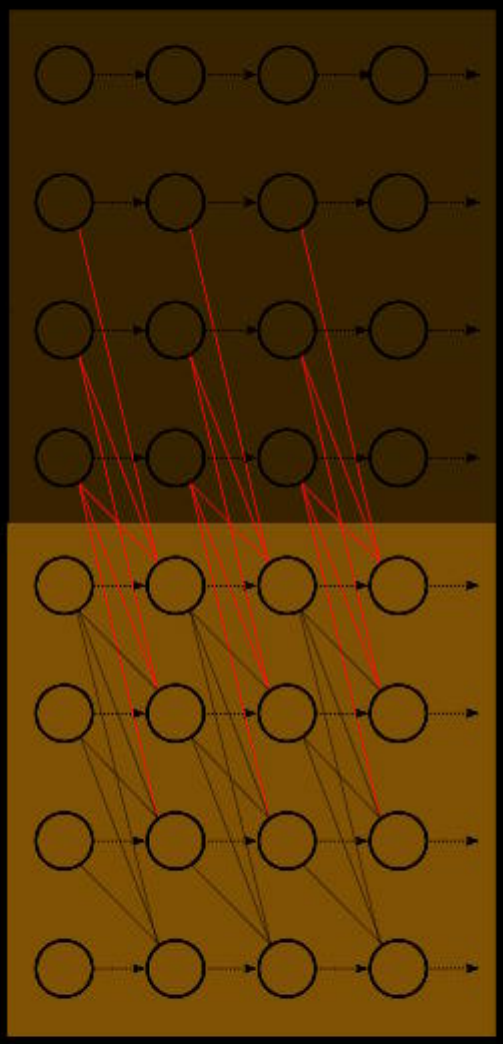
[35] S. Wang, X. Cheng, Y. Li, M. Wu, and Y. Zhao. Image-based promoter prediction: a promoter prediction method based on evolutionarily generated patterns. 8(1):17695.

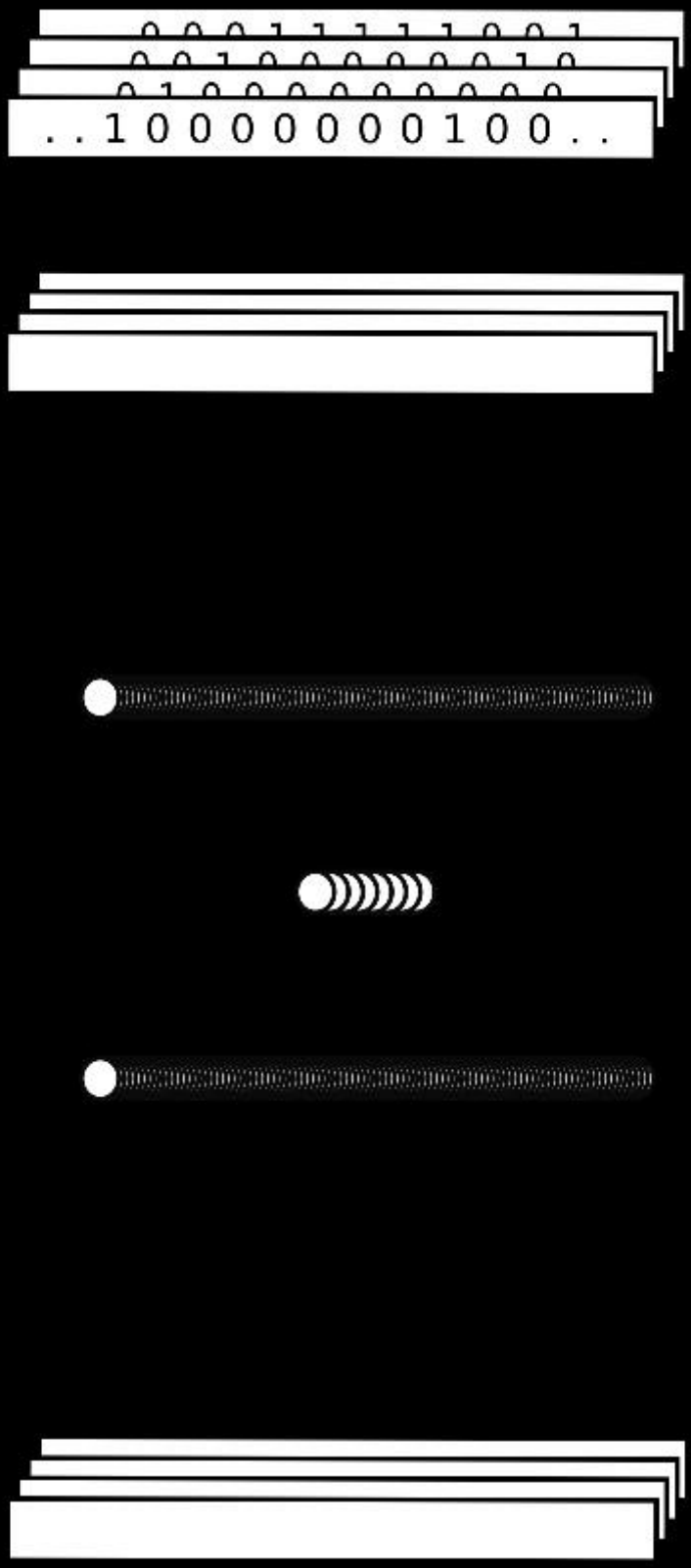
[36] B. Yan, M. Boitano, T. A. Clark, and L. Ettwiller. SMRT-cappable-seq reveals complex operon variants in bacteria. 9(1):1–11.

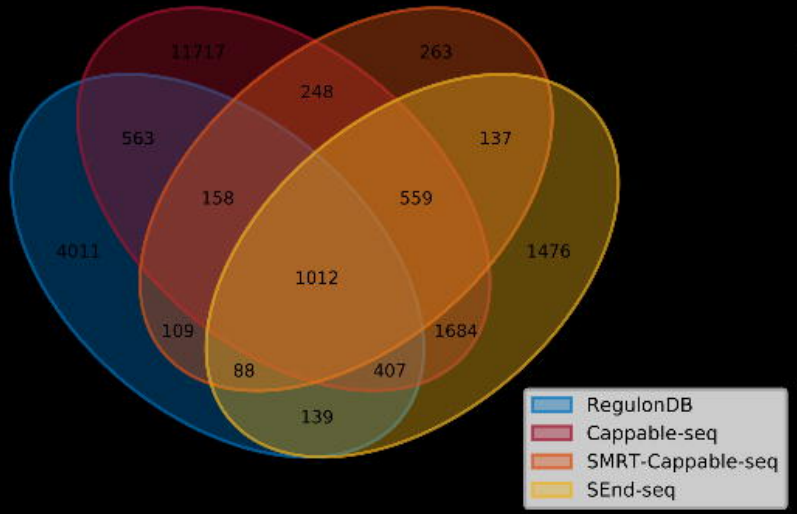
[37] J. Zhou, C. L. Theesfeld, K. Yao, K. M. Chen, A. K. Wong, and O. G. Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*, 50(8):1171–1179, Aug. 2018.

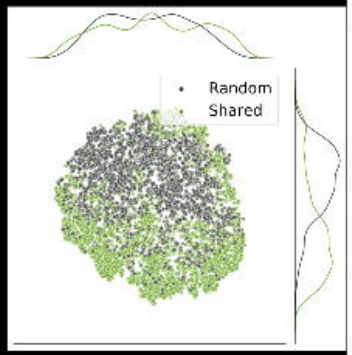
[38] J. Zhu and M. Q. Zhang. SCPD: a promoter database of the yeast saccharomyces cerevisiae. 15(7):607–611.

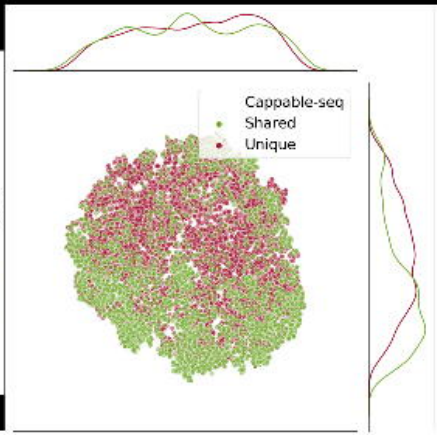
[39] T. Łozinski, W. T. Markiewicz, T. K. Wyrzykiewicz, and K. L. Wierzchowski. Effect of the sequence-dependent structure of the 17 bp AT spacer on the strength of consensus-like *E.coli* promoters *in vivo*. 17(10):3855–3863.

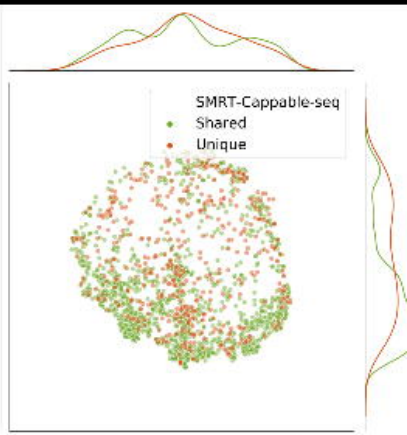


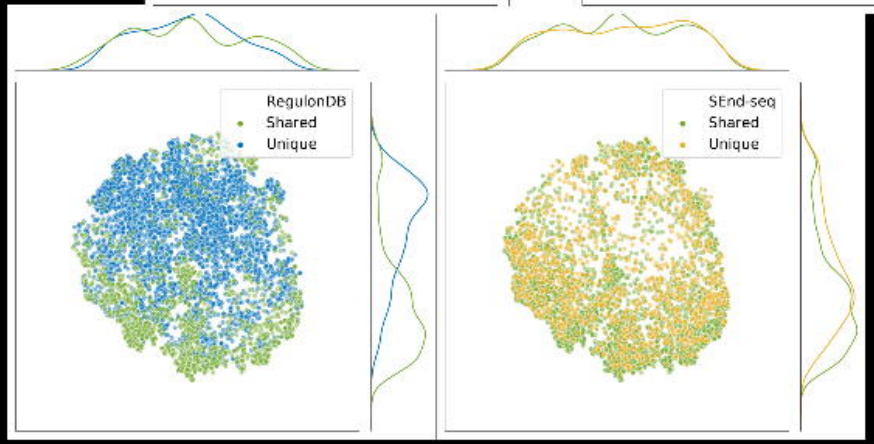

. . 1 0 0 0 0 0 0 0 0 1 0 0 . .

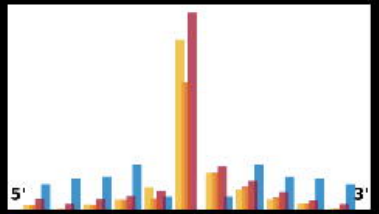

11717
263
248
563
137
158
559
4011
1476
1012
109
1684
88
407
139
RegulonDB
Cappable-seq
SMRT-Cappable-seq
SEnd-seq

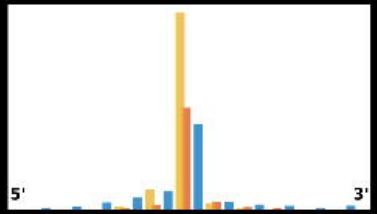

Random
Shared

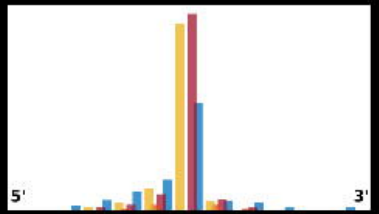

Cappable-seq
Shared
Unique

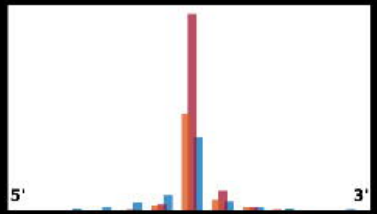

SMRT-Cappable-seq
Shared
Unique

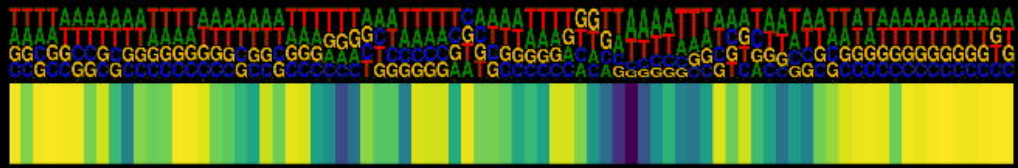

RegulonDB
Shared
Unique
SEnd-seq
Shared
Unique

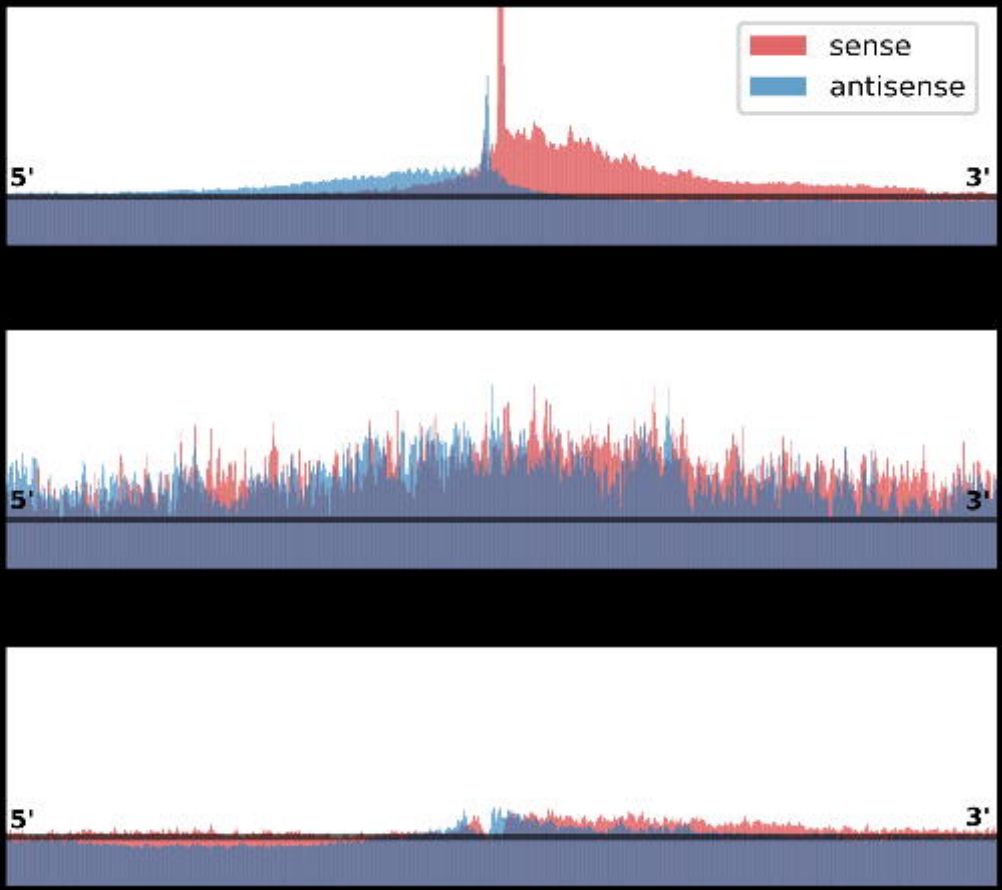
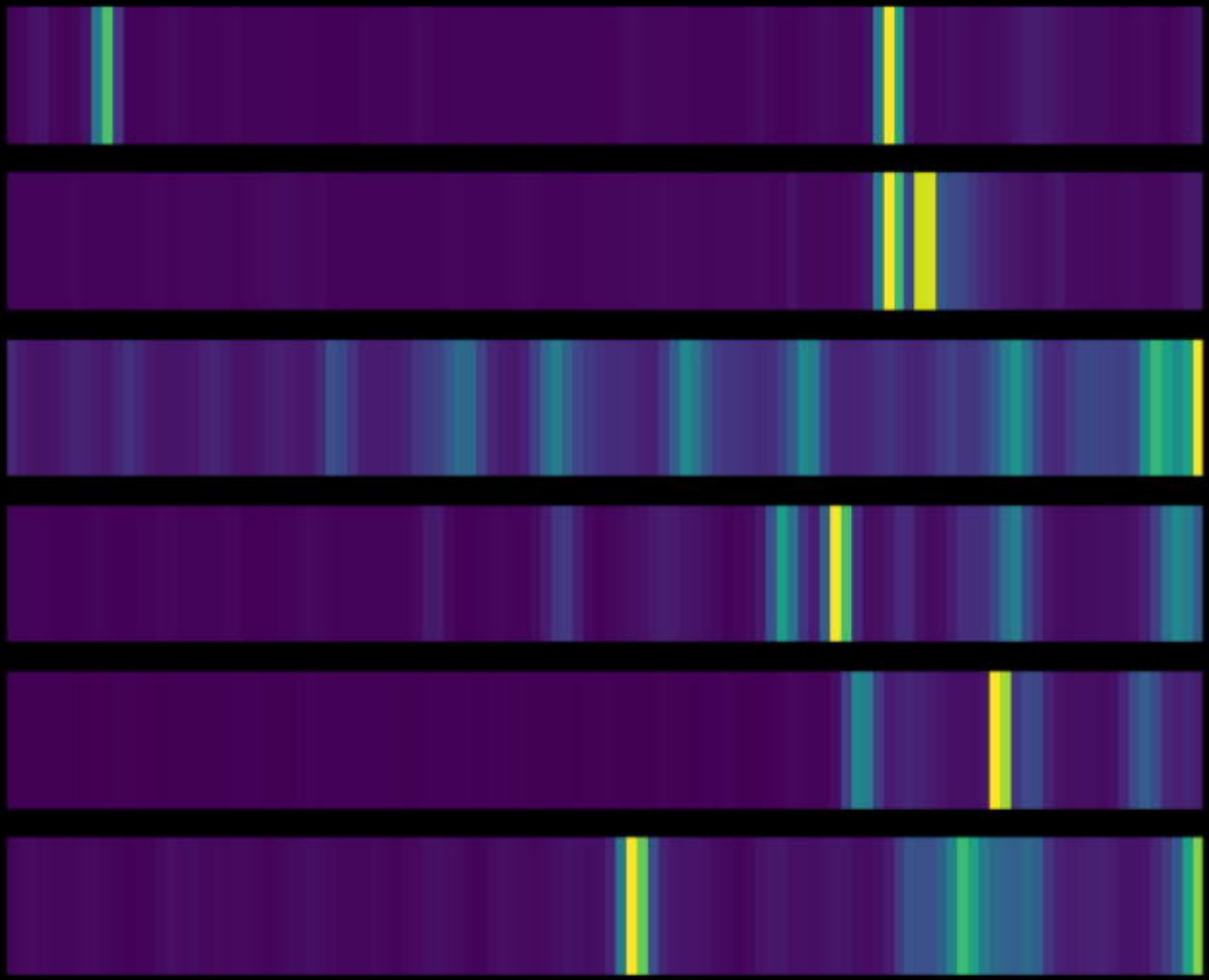

5'
3'


5'
3'


5'
3'


5'
3'




sense
antisense
5'
3'
5'
3'
5'
3'