AN ABSTRACT OF THE THESIS OF

Taj H. Morton for the degree of Master of Science in Computer Science presented on November 7, 2014.

Title: Prediction of Gene Transcription Start Sites and Initiation Patterns from DNA Sequence Content.

Abstract approved:

_____

Molly Megraw, Weng-Keen Wong

The computational identification of gene Transcription Start Sites (TSSs) can provide insights into the regulation and function of genes without performing expensive experiments, particularly in organisms with incomplete annotations. High-resolution general-purpose TSS prediction remains a challenging problem, with little recent progress on the identification and differentiation of TSSs which are arranged in different spatial patterns along the chromosome.

In this work, we present TIPR, a sequence-based machine learning model which identifies TSSs with high accuracy and resolution for multiple spatial distribution patterns along the genome, including broadly distributed TSS patterns which have previously been difficult to characterize. TIPR predicts not only the locations of TSSs,

but also the expected spatial initiation pattern each TSS will form along the chromosome—a novel capability for TSS prediction algorithms. As spatial initiation patterns are associated with spatiotemporal expression patterns and gene function, this capability has the potential to improve gene annotations and our understanding of the regulation of transcription initiation. The high nucleotide-resolution of this model locates TSSs within 10 nucleotides or less on average.

Prediction of Gene Transcription Start Sites and Initiation Patterns from DNA Sequence Content

by
Taj H. Morton

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented November 7, 2014
Commencement June 2015

Master of Science thesis of Taj H. Morton presented on November 7, 2014.

APPROVED:

Co-Major Professor, representing Computer Science

Co-Major Professor, representing Computer Science

Director of the School of Electrical Engineering and Computer Science

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Taj H. Morton, Author

ACKNOWLEDGEMENTS

I would like to thank everyone who has assisted, supported, and mentored me in this project, especially my advisors Dr. Molly Megraw and Dr. Weng-Keen Wong. I would also like to thank my committee for their time and guidance: Dr. David Hendrix and Dr. Rob Holman. Finally, I would like to thank my research group—Mitra Ansariola, Stephanie Bollmann, Jason Cumbie, Sergei Filichkin, Maria Ivanchenko, and Spencer Kisler—for all their advice and camaraderie which has made this work possible.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1 Introduction

Transcription Start Sites (TSSs) and their associated promoter regions play a critical role in the transcription of genes. However, the mechanisms by which transcription is initiated at specific genomic locations is still not fully understood, including how the spatial distribution of TSSs is defined, how promoter architecture influences this spatial pattern, and how genes lacking canonical elements within the core promoter are transcribed. The advent of high-throughput TSS sequencing protocols such as CAGE and PEAT have transformed the field of promoter analysis, providing genome-wide nucleotide-resolution information on TSS usage (Carninci et al., 2005; Ni et al., 2010). One important goal in this field is the identification of TSS locations when TSS-Seq data is unavailable. While start codons are easily identified, the length of the 5' UTR upstream of the first exon varies from gene to gene and even between transcripts of the same gene, yielding different mRNA products. Several studies have taken computational approaches to TSS identification, building machine learning models which predict the location of TSSs from the surrounding sequence content with varying degrees of success and resolution, ranging from the prediction of individual nucleotides to regions up to 500 nt wide (Abeel et al., 2009; Boer et al., 2014; Knudsen, 1999; Megraw et al., 2009; Morton et al., 2014; Ohler et al., 2000; Sonnenburg et al., 2006; Zhao et al., 2007).

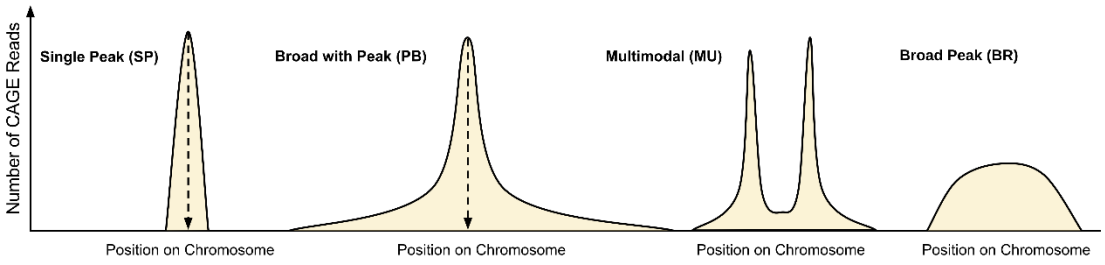


Figure 1: Common TSR Initiation Patterns

The four Transcription Start Region initiation patterns introduced by Carninci et al. (2006), as identified in genome-wide mouse and human CAGE studies. In this work, we focus on the Single Peak and Broad Peak initiation patterns, also identified in *Drosophila melanogaster* (Ni et al., 2010) and *Arabidopsis thaliana* (Morton et al., 2014).

The mRNA products produced during the transcription of genes typically do not all initiate at a single genomic location. Instead, transcription initiates upstream of the gene's start codon in a region that can range from vary narrow (2—3 nt) to wide (upwards of 50 nt or more), forming a collection of individual TSSs known as a TSS cluster or TSR (Transcription Start Region) (Carninci et al., 2006; Ni et al., 2010; Rach et al., 2009). TSS clusters can be grouped by the width and distribution of individual TSSs that define the cluster (Figure 1). In this study, we focus on the Single Peak (or Narrow Peak) and Broad Peak (or Weak Peak) patterns defined in previous TSS-Seq studies (Carninci et al., 2006; Ni et al., 2010). Previous studies have shown that different initiation patterns are associated with different types of genes, tissues, and regulatory mechanisms such as Transcription Factors (TFs) and CpG islands (Morton et al., 2014; Ohler and Wassarman, 2010; Sandelin et al., 2007). While there

has been success in the identification of Narrow Peak initiation patterns (Megraw et al., 2009), it has remained unclear whether other initiation patterns can be predicted from sequence content alone at the same nucleotide-level resolution. Models incorporating additional data types such as histone modifications have had success in the prediction of these less well-defined patterns (Rach et al., 2011), though prediction of broadly distributed patterns is still clearly a greater challenge. An analysis of 17 TSS prediction models found that these broad patterns could be predicted with low resolution (500 nt) from sequence content alone, but did not explore nucleotide-resolution models (Abeel et al., 2009).

In this work we present a machine learning model capable of predicting TSSs of multiple initiation patterns with high performance and positional resolution, while also suggesting the probable initiation pattern the TSS cluster would form along the chromosome. The TIPR model utilizes features derived from sequence content and TF binding affinity to predict the probability of transcription initiation at an individual nucleotide. Because this model provides nucleotide resolution and initiation pattern prediction, the model can be used to infer answers to a wide variety of topics, including a better understanding of promoter architecture, improved gene finding and annotations, identification of TFs which could be involved in the regulation of genes, and positional information guiding wet-laboratory experiments. We evaluate the TIPR model using publicly available high-throughput TSS-Seq datasets from mouse (Carninci et al., 2006) and *Arabidopsis thaliana* (Morton et al., 2014). Our model

performs well (AUROC 0.99, AUPRC 0.82), demonstrating that TIPR can successfully predict TSSs across multiple organisms and tissue types. TIPR uses only sequence information, and is therefore applicable in cases where TSS-Seq data is not yet available.

2 Literature Review

Previous studies have constructed machine learning models which predict TSSs with varying degrees of accuracy and resolution. The first TSS prediction models grew out of gene finding programs. At the most basic level, these tools searched for simple sequence content enrichments or well-known binding elements, such as the TATA box and CpG islands. Many early TSS predictors suffered from issues with high numbers of false positives (Fickett and Hatzigeorgiou, 1997), limiting their use.

PromoterScan (Prestridge, 1995) was one of the first TSS predictors to utilize a large collection of Transcription Factor Binding Site (TFBS) elements as predictors. Because most TFBSs were not well-characterized at this time, PromoterScan used subsequence matching to score the number of detected promoters within a sequence of interest. The importance of a TFBS was calculated by computing a density ratio comparing the presence of the TFBS in positive (TSS-containing sequences) and non-transcribed sequences. These ratios are used to form a Promoter Recognition Profile, essentially ranking TFBSs by how discriminative they are between TSS and non-TSS sequences. Predictions were made by scoring a sequence based on the presence of TFBSs within a 250 nt window upstream of the TSS, weighted by their Promoter Recognition Profile score. In addition, the well-characterized TATA box was considered separately, using a published PWM derived from 502 Pol-II transcribed regions (Bucher, 1990) within the 3’-most 50 nt of the sequence. PromoterScan performed well compared to other TSS predictors at the time, however the resolution

of the model was limited both by the coarse windowing procedure and data available at the time.

Promoter2.0 (Knudsen, 1999) took inspiration from neural networks and genetic algorithms to identify sub-patterns within DNA sequence and discriminate between promoter-containing and non-promoter genomic regions. Neural networks are used to model TFBSs as *k*-mers, scoring a region of DNA by measuring the highest activity of the output neuron within the window. Each *k*-mer is represented as a binary vector of length 4*k* and used as input for the neural network. Neural networks are trained individually by changing the value of a randomly selected weight by a random amount during each generation and evaluated by computing the correlation coefficient and SSE of classifying training data using the current model. The study evaluated neural networks with random initial weights, as well as networks which were designed to model the binding sites of 4 TFs, TATA box, the cap site, CCAAT box, and the GC box. Promoter2.0 was evaluated on a variety of promoter sequences available at the time, including the Bucher database of promoters (Bucher and Trifonov, 1986) and the complete adenovirus 2 genome. The Promoter2.0 model scores windows of sequence, assigning them scores corresponding to how promoter-like the sequence is. Therefore, the model's resolution is limited by the width of the sequence windows used in training and testing. In testing, the model predicted several TSSs within 161 nt of the true TSS.

More advanced predictors were developed in the early 2000s, taking advantage of the availability of larger data sets and more complex modeling techniques. McPromoter (Ohler et al., 2000) uses a generalized form of hidden Markov models (HMMs) known as a stochastic segment model (SSM). Segment models can be defined for different regions of the promoter, such as the region upstream of the TSS, the core promoter, and the region directly downstream of the TSS. These regions can be further separated as desired for more complex modeling, such as separating the core promoter into regions which are bound by different elements of the transcription preinitiation complex like TATA and Initiator. Segment models had been previously used in gene finding models, segmenting genes into regions such as the start codon, introns, exons, splice sites, and so on. However, unlike genes, promoters do not contain such universal, well-defined segments. Because there is no clear universal promoter structure (combination of TFBSs), we know neither the number of segments nor the positions where they would fall. Each segment contains an output distribution model, used to generate the most probable sequence given a segment. McPromoter uses fourth-order Markov chains as the output distribution of each segment. A non-promoter model was created with a mixture distribution of two Markov chains, one trained on coding sequences and the other on intronic sequences. Unlike the earlier models discussed above, McPromoter scores smaller windows of sequence (on the order of 10-50 nt), producing a much higher resolution prediction signal. However,

Megraw et al. (2009) reports that while McPromoter can technically achieve high resolution, predicted TSSs are often 80 nt or further from the true TSS.

ARTS (Sonnenburg et al., 2006) takes a machine-learning approach to promoter prediction, building a support vector machine (SVM)-based model. The ARTS model uses combinations of SVM kernels which capture sequence similarity between training examples, TFBS sequences located in the promoter region, and the 3D structure of the DNA molecule. Specifically, four kernels are utilized by ARTS, each capturing a different aspect of TSS complexity. A WD_S (Weighted Degree Kernel with Shifts) kernel (Rätsch et al., 2005) captures similarities between sequences which have been shifted relative to a TSS, but are otherwise very similar in content and order. This kernel gives some flexibility for the position of elements relative to the TSS. A second kernel captures the presence of TFBSs without penalizing them for location or order. This spectrum kernel (Leslie et al., 2002) measures the over- or under-representation of a TFBS within the promoter. A separate spectrum kernel is applied to the sequence downstream of the TSS in the gene's 5' UTR, coding, and intronic regions due to the difference in sequence and element composition compared to the promoter region. Two linear kernels are used to incorporate the DNA molecule's 3D structure, capturing the twisting angles and stacking energies of windows of sequence. The ARTS model can be efficiently trained and evaluated on a genome-wide scale and predicts regions containing TSSs with high accuracy. In an analysis of 17 TSS prediction models, ARTS out-performed all other models on every

metric (Abeel et al., 2009). In a comparison of 4 models, ARTS correctly predicts most TSSs with few false positives (Megraw et al., 2009). However, the model under-performs at identifying the precise genomic location of TSSs, with many predicted TSS locations falling between 60 – 80 nt from the true TSS.

CoreBoost (Zhao et al., 2007) utilized a decision-stump based model for TSS prediction, and focused on making predictions with high resolution over smaller genomic regions. CoreBoost considered CpG-related and non-CpG-related promoters separately, building separate models for each promoter class. This approach was taken to focus on the more difficult non-CpG-related class of promoters, similar to the separate initiation pattern classifiers we have constructed in this work. CoreBoost models transcription as a hierarchical system, allowing for more flexible decision boundaries. Models included core promoter elements (such as TATA and Inr), TFBSs from TRANSFAC, sequence content enrichment of dimers, third-order Markov models, and structural properties of DNA molecules. The combination of all of these features was critical to CoreBoost's good performance. CoreBoost is implemented using LogitBoost with decision stumps, with separate families of classifiers trained for CpG and non-CpG promoters. In addition, separate class labels are assigned to the upstream and downstream regions surrounding a TSS due to their differing structure and content. This allows classifiers to pick the most discriminating features for their specific class and contributed to a boost in CoreBoost performance. CoreBoost is designed for the identification of TSSs with high resolution—identifying their

locations very accurately within a region of the genome. However, the authors recommend using another method to focus CoreBoost's predictions within a 2.4 kb (or smaller) region using other data sources, such as ChIP-chip, ESTs, mRNAs, or gene finding programs, and note that CoreBoost's performance increased substantially when Chip-CHIP data was used to identify promoter-containing regions.

As the volume and resolution of available high-throughput TSS sequencing data increased, it became possible to train and evaluate models more accurately. The TSS prediction models reviewed here were trained and evaluated using older and lower-throughput protocols, such as data curated by the DBTSS project (Wakaguri et al., 2008). Most TSSs in these datasets were identified from low-resolution sources such as cDNAs and ESTs. New high-throughput TSS-Seq protocols like CAGE and PEAT have revolutionized the availability, quality, and resolution of TSS data. These protocols have been used to generate genome-wide, nucleotide-resolution TSS datasets in multiple species, including human and mouse (Carninci et al., 2006), *Drosophila melanogaster* (Ni et al., 2010), and *Arabidopsis thaliana* (Morton et al., 2014). The availability of this data has inspired several new TSS models (Boer et al., 2014; Megraw et al., 2009; Morton et al., 2014) and re-analysis of existing methods (Abeel et al., 2009), trained and tested with higher-resolution data.

S-Peaker (Megraw et al., 2009) was the first TSS prediction model to take advantage of genome-wide TSS-Seq datasets for TSS prediction. Using L1-regularized logistic regression and TFBS PWMs from TRANSFAC (Wingender, 2008), S-Peaker is an

interpretable, high-resolution model which is capable of predicting Narrow Peak TSSs in mouse. When compared directly, S-Peaker identified the locations of TSSs with higher accuracy than ARTS, CorePromoter, or McPromoter achieved on the same test set. However, S-Peaker was limited to the prediction of TSSs with Narrow Peak initiation patterns, while the other methods were capable of predicting TSSs of any type.

In this work, we build on S-Peaker and 3PEAT (Morton et al., 2014) to build a model capable of predicting both Narrow Peak and Broad Peak initiation patterns with a single model, while maintaining the resolution and preciseness of the S-Peaker model. We also investigate the impact of L1-regularized regression and explore alternative regularization techniques which could produce more interpretable models, while still maintaining the same prediction performance. Previous models have primarily focused on the latter, while the former was often considered a secondary priority. An L1-regularized logistic regression model is easily interpretable: the most informative predictors are assigned the highest weights, while uninformative features are removed from the model entirely. Other modeling techniques such as SVMs and HMMs can define non-linear decision boundaries, but are more difficult to interpret owing to their higher-order and more flexible nature. The TIPR model achieves both of these goals: TSSs are predicted with high accuracy and resolution while being constructed using a simple combination of multiple logistic regression models. At the same time, TIPR

achieves better performance (in both AUROC and AUPRC) than the older S-Peaker model.

3 Methods

3.1 Overview of TIPR Pipeline

Our TSS prediction pipeline begins with the creation of a dataset containing the genomic locations of TSSs identified by high-throughput TSS-Seq protocols. In this analysis we have restricted our model to the prediction of protein coding genes.. While other products such as miRNAs, snoRNAs, and lncRNAs are also Pol-II transcribed, the promoter architecture of these products are not as well-understood as protein-coding mRNAs, and in some cases may utilize unique promoter elements and regulatory programs (Alam et al., 2014). Therefore, we restrict our analysis to TSSs which are located no further than 500 nt upstream of a protein-coding gene's annotated 5' UTR. TSS tag clusters (spatially grouped TSS-Seq reads) are next filtered by read count, ensuring that only commonly-transcribed TSSs are used to build the model. After filtering, TSS tag clusters are grouped by initiation pattern (Single Peak and Broad Peak, Figure 1) into individual datasets. Finally, the mode of each tag cluster (the nucleotide where transcription most frequently initiates within the cluster) is determined and used as a single, putative genomic location for the tag cluster.

After the set of TSS tag clusters are created, 5 KB of genomic sequence is extracted upstream and downstream of each tag cluster mode. The sequences are converted into numerical features representing the presence of general transcription factor binding

sites (including TFBSs and TATA-binding protein associated sites) in regions where they are likely to be functional and involved in recruitment of transcription machinery, following the procedure described in Megraw et al. (2009). In this work, we use TFBS as a general term for all vertebrate binding sequences described by the TRANSFAC database (Wingender, 2008). These include both transcription factor binding sites and TATA-binding protein associated (TAF) site sequences. In addition to positive examples (locations where transcription initiates), negative examples (locations with no evidence of transcription initiation) are selected by randomly choosing genetic sequence from genic, intergenic, and promoter-proximal regions.

Four different binary logistic regression classification models are constructed from the training dataset, shown in Table 1. Models are constructed using a modified version of the l1_logreg package, an implementation of the interior-point method for L1-regularized logistic regression (Koh et al., 2007). Cross-validation is used to select the TIRP model parameters. The optimal L1 penalty parameter λ is chosen by finding the λ values which yield the highest AUROC in each validation partition, and computing the average of these values. A second parameter *d* used by the SP vs BR model is selected on a secondary held-out validation partition by F1 score. After parameter selection, final models of each type are constructed using the entire training dataset with the optimal λ parameter.

Finally, the model is evaluated by classifying examples from a held-out test set, comprised of 20% of all examples in each dataset (including negative examples

described above) and an additional 100,000 negative examples drawn randomly from the entire genome. Each model is used to classify every test-set example, producing a total of 4 probability values per example. The SP vs BR classifier is used to select the appropriate TSS vs NoTSS classifier for a given example, based on the *d* parameter calculated during training and cross-validation. This process can be applied on a genomic scale by repeating this prediction process at every nucleotide in the region of interest, producing a signal along the chromosome representing the probability of transcription initiation at each nucleotide. Figure 2 shows a flow chart summarizing our TSS prediction data preparation pipeline and classification process.

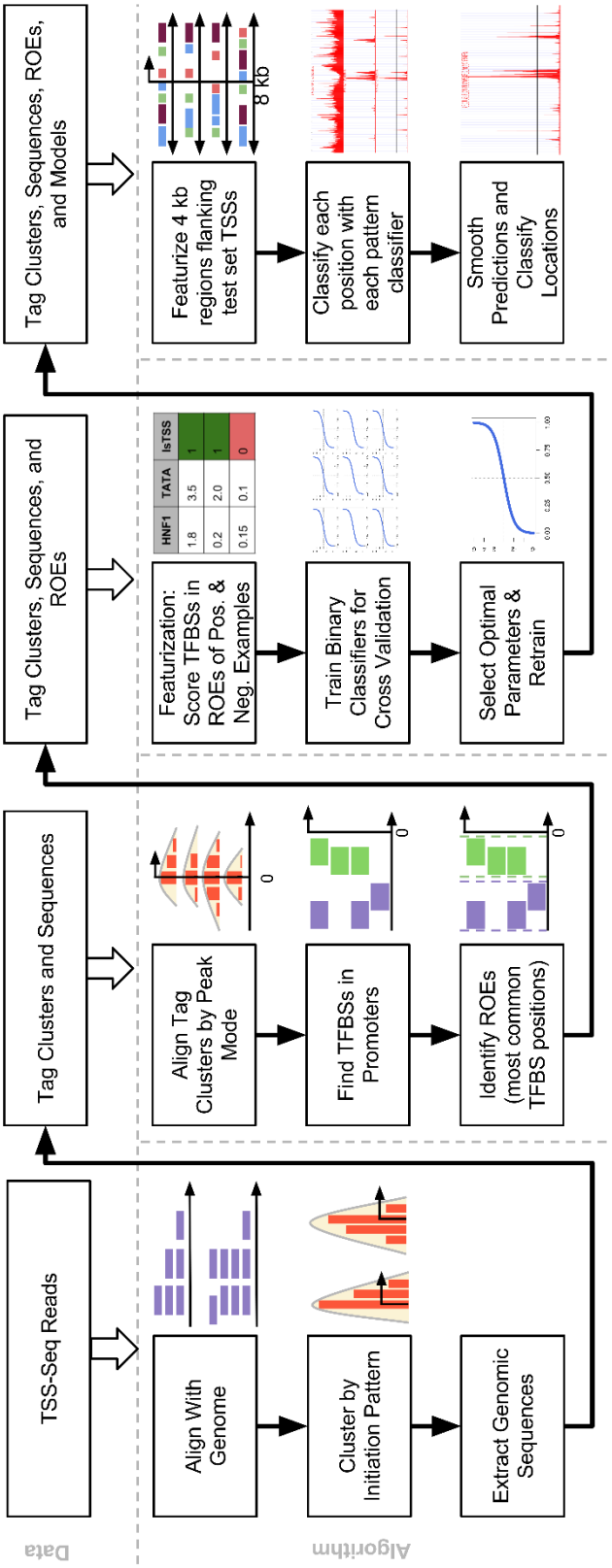


Figure 2: Flowchart of the TIPR pipeline

Table 1: List of binary classification models trained and tested in this study

Model Name	***Class 1***	***Class 2***
SP vs NO	Single Peak TSSs	Negative (Non-TSS) Genomic Locations
BR vs NO	Broad Peak TSSs	Negative (Non-TSS) Genomic Locations
SP vs BR	Single Peak TSSs	Broad Peak TSS
ALL	SP and BR TSSs	Negative (Non-TSS) Genomic Locations

3.2 Identification of TFBS Regions of Enrichment

In this study, TFBSs are characterized by experimentally supported Positional Weight Matrices curated by the TRANSFAC project (Wingender, 2008) which approximate the affinities of many Transcription Factors for potential DNA binding sequences. Because TFBSs are often short, degenerate sequences, they occur frequently throughout the genome for many TFs. Even if we assume that TF binding does occur at every TFBS location that occurs in the genome, a majority of this binding almost certainly does not lead to transcription. For example, the TATA box TFBS is typically located in a window 25 – 35 bp upstream of the TSS, where it binds to the TFIID protein, forming a multi-protein complex which binds to the Pol-II complex and initiates transcription. If a TATA box binding site is observed hundreds of base-pairs upstream from a TSS, it is unlikely that this TATA site is involved in the transcription of this TSS. Therefore, as part of our training process, we

computationally identify regions of the promoter in which each TFBS in our dataset is likely to be functional. This procedure specifically focuses our model on TFBSs located in regions of the promoter where they are likely to be involved in transcription, as opposed to including every TFBS located near a TSS, regardless of location. We call these locations "Regions of Enrichment," as they are regions positioned relative to TSSs in which a TFBS is significantly enriched compared to the promoter background sequence distribution. Our machine learning analysis is restricted to TFBSs which fall within these regions. This technique has two major advantages. Firstly, it serves as a feature reduction technique, enabling faster model training and testing. Secondly, it allows the model to identify features which are more likely to be biologically relevant.

To identify these Regions of Enrichment, we consider all TSS tag clusters grouped by TSS initiation pattern. TFBS PWMs are scanned along regions 2kb upstream and downstream of the TSS, computing the log-likelihood score of the TFBS at every nucleotide compared to the promoter background distribution. These scores are combined and averaged into a single score at each nucleotide. Starting from the highest scoring nucleotide within 100 nt of the TSS, the ROE is expanded left and right until the log-likelihood score falls below the average TFBS score of the promoter (within 2kb of the TSS) for at least 5 nt. This region represents the most-common positions in which TFBSs for a particular TF occur relative to the TSS. During featurization and prediction, only TFBSs which fall within the ROE are considered by the TIPR model. Figure 3 shows a diagram of this process.

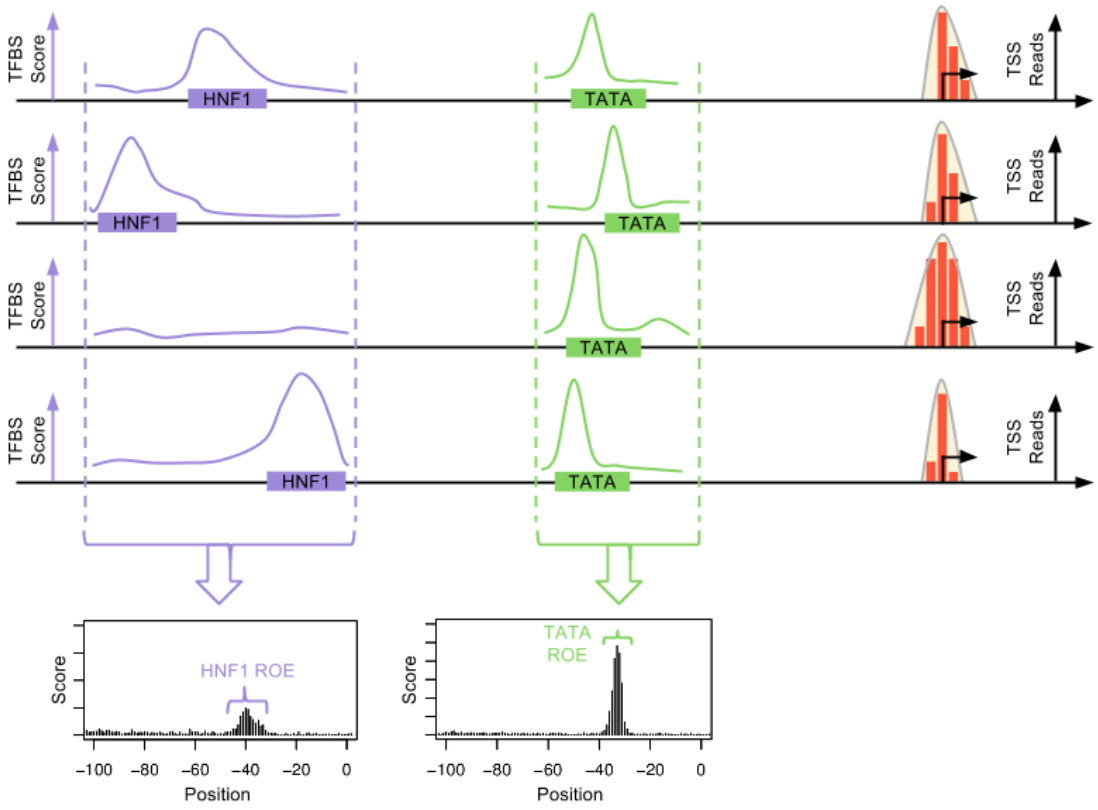


Figure 3: Diagram of ROE Identification

Diagram of how ROEs are identified from raw TSS-Seq reads. Individual reads (tags) are aligned (orange bars) and clustered together by their location (each row represents a TSS tag cluster). The most highly expressed nucleotide within each cluster is considered to be the putative TSS location and assigned a relative position of 0. The DNA sequence surrounding each tag cluster is extracted and TFBSs are identified and scored by log-likelihood score (purple and green). These scores are averaged across all tag clusters (bottom plots) and the region most enriched for a particular TFBS is selected as the TF's ROE (dashed lines).

3.3 Featurization and Model Construction

After TSS tag clusters have been identified from the TSS-Seq data and Regions of Enrichment have been defined, we convert the DNA sequence surrounding TSSs into

numerical features for the purposes of model training and evaluation (Figure 4). These numerical features characterize the presence of TFBSs within ROEs. Within the ROE of a given TFBS, the log-likelihood score of the TFBS PWM (with respect to the promoter background distribution) is calculated at every nucleotide within the ROE and summed together to produce a single numerical score. This score is large when the promoter sequence of a TSS closely resembles the TFBS and small when the ROE does not contain any sequences which closely match the TFBS of the given TF. In order to increase the resolution of the features and allow the model to select the most informative locations, each ROE is split into 7 sub-regions, 5 central overlapping windows and 2 flanking windows (Megraw et al., 2009, Figure 3). In addition, sequence enrichments for GC, GA, and CA dinucleotides surrounding the TSS are computed and included as features. We also considered a higher-order model which was composed of the above features plus all pairwise interactions between features. Pairwise interactions were computed by multiplying each feature by every other in the feature model, producing a model containing a total of n^2 features.

The ROEs used to construct the ALL model were selected by combining the SP and BR datasets together before performing ROE selection. For the sub-models of the MSC classifier, ROE selection was performed on individual initiation pattern datasets, and the resulting ROEs from each dataset were combined together. This was done because initiation patterns seem to have distinct promoter architectures and differing preferred locations for TFBSs relative to the TSS.

Negative training and testing examples are featurized as above, but rather than using TSS-proximal sequences, they are instead composed of randomly selected genomic locations at which there is no evidence of transcription initiation. In order to create a high-resolution model that performs with high sensitivity and specificity, the model must differentiate between true TSSs and nearby sequences which are not transcribed, but which have similar sequence content. To ensure the model training and testing sets support this goal, we select 20 negative examples for every 1 positive TSS example which are drawn from genomic locations located 200 to 2000 nt upstream of the TSS. In addition, we also draw 1 negative examples from exonic and intergenic regions for every positive example in the training set. Finally, an additional 100,000 negative examples are drawn randomly from the entire genome and used for testing. In future work, selecting the 100,000 negative examples only from regions where transcription initiation of any type is not known to occur may improve the model's performance in testing. However, given the size of mammalian genomes, it is unlikely that a significant number of TSSs (if any) were included in this test set. In addition, because these randomly-selected examples are not used in model training or parameter selection, false negatives within this set don't affect the final model, only underestimate the sensitivity/recall reported in testing.

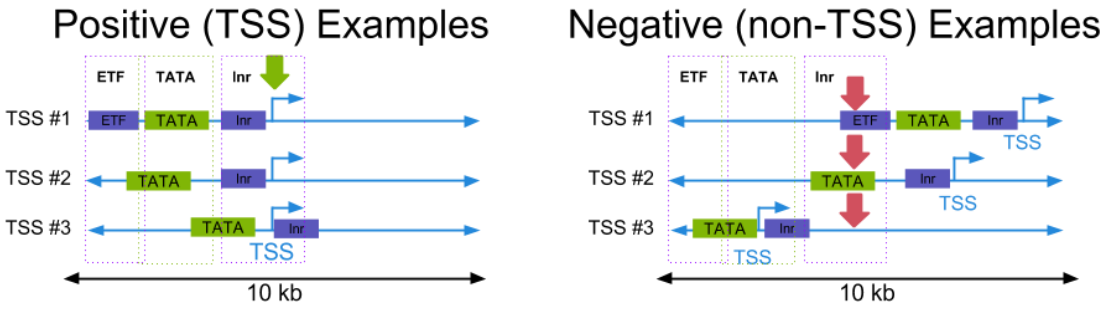


Figure 4: Diagram of sequence featurization process

The DNA sequence surrounding experimentally identified TSSs are extracted (labeled TSS, green arrow) and the presence of absence of TFBSs within their ROEs are scored. The ROEs identified for each TF are shown in dotted lines. Red arrows denote the positions of randomly-selected negative examples where no evidence of transcription was supported by the TSS-Seq dataset.

3.3.1 Flanking Features

The featurization method described above assumes that TFBSs are located relative to the mode of the TSS tag cluster. While this site is the most commonly transcribed location within the cluster, many other sites surrounding the mode are also transcribed, especially within the broad initiation pattern. We hypothesized that the presence of TFBSs located nearby to (but outside of) the TSS's ROEs could also be informative to the model, especially for discriminating between different initiation patterns. Therefore, we created another set of features we called "flanking features." These are additional features added to models which characterize the sequence content in the vicinity of the TSS mode, capturing TFBS scores upstream and downstream of the TSS mode's ROE for each TF. Flanking features are computed by essentially

shifting the “0” position which is used to compute the location of ROEs relative to the “0” position. For example, the TATA ROE is a region 25 – 35 nt upstream of the TSS. When a flanking feature 10 nt upstream of the TSS mode is computed, the TATA ROE region is shifted to fall 35 – 45 nt upstream of the TSS mode. We tested a variety of flanking widths. Positions up to 200 nt upstream and downstream of the TSS mode were included in the model dataset, with 1 feature window computed every 10 nt, yielding a total of 40 additional features (20 upstream and 20 downstream of the TSS).

3.3.2 Featurization of Conservation Information

Sequence conservation has been shown to be a useful predictor of the functionality of DNA sequences, including TFBSs (Jin et al., 2006). When a region of DNA is conserved, it is thought that this region must be under evolutionary pressure, where mutations are harmful and do not survive. Conservation scores represent the stability of a nucleotide over evolutionary time. The PhastCons score (Siepel et al., 2005) compares the sequence similarity of genes and orthologs of these genes in other related species. In this study, we investigated multiple methods for including sequence conservation information within the model.

The PhastCons scores from the most recent mouse genome (mm10) were used to calculate the conservation of the promoter region of each TSS. In this study, we considered the ROE of each TFBS separately, reasoning that even though within a gene’s promoter region, important TFBSs would likely be conserved, but sequence between the TFBSs would not necessarily be under evolutionary pressure. For each

ROE, a score was computed by summing all PhastCons scores within the ROE. Missing scores were presumed to be 0. Models using this final score feature as well as models using sub-score features (matching the ROE sub-windows) were considered. Two different methods of combining these resulting features with the model were considered. In the first model, these conservation scores were included directly in the model as features. In the second model, conservation scores were multiplied by the TFBS's cumulative log-likelihood score, creating a combination feature influenced by both the score of the TFBS and how well-conserved the site is. These models included both the original log-likelihood score feature along with the combined log-likelihood and conservation feature.

This representation of conservation is not without issues, however. For example, conservation scores are typically computed by aligning segments of genomes to discover genes which are common across multiple organisms. However, TFBSs are short, degenerate sequences which are found throughout the entire genome, so the alignment of promoter regions requires special attention. A naïve approach is to align all gene orthologs by their start codon and make no attempt to align the 5' UTR or promoter region of each ortholog separate. This causes multiple issues, as 5' UTRs are variable in length (especially across species), and exact TSSs are not well-characterized in many species. Even assuming that the TSSs of orthologs could be identified and used as a common reference point, the locations of functional TFBSs within promoters are not fixed, especially across species. An analysis by Kunarso et

al. (2010) showed that the occupancy profiles of the transcription factors OCT4 and NANOG in embryonic stem cells are significantly different between mouse and human. While genes which were regulated by OCT4 were enriched for OCT4 and NANOG in both mouse and human, the locations of these binding sites were often not conserved (Kunarso et al., 2010; Villar et al., 2014). More concretely, we can imagine a TF with a functional in a region of [-20, -10] in one species, but is functional in the region [-100, -90] in another. A more complex modeling approach would be to compute ROEs across all related species and calculate conservation (or TFBS log-likelihood) scores within the appropriate ROE for each species.

3.4 Model Construction

After featurization, the TIPR model is constructed by training the 4 models listed in Table 1 independently, trained on 80% of the dataset. 80% of each cross-validation fold is used for model training, 10% for regularization parameter (λ) selection, and the remaining 10% for the SP vs BR cutoff threshold (d) parameter selection. The regularization parameter λ is selected by choosing the value which provides the highest AUROC on the validation partition of each fold. The cutoff parameter d is selected by choosing the value which optimizes the classifier's F1 score over the held-out partition. The optimal λ of each fold is used in the classification of these examples. The F1 score is the harmonic mean of precision and recall; it is used to select a classifier that is optimized to predict both SP and BR initiation patterns successfully.

After all parameters are selected and models have been built using the full 80% of the training data, the held-out testing data is classified by each model independently. After classification by the binary models (SP vs No and BR vs No), a two-stage classifier is used to produce the final call. When predicting an individual example, first the SP vs BR prediction is examined, determining if this nucleotide more closely resembles a SP or BR initiation pattern. If the SP vs BR prediction predicts SP (above the SP_vs_BR threshold *d*), the prediction of the SP vs No classifier is used to predict the final class label. Conversely, the BR vs No classifier is used to predict the final class label when the SP vs BR model predicts the location resembles a BR initiation pattern. We call this classifier the MSC (multi-stage classifier) model, as it applies a hierarchical procedure for determining the appropriate classification models. This allows for more flexibility in the selection of probability cutoff thresholds.

3.5 Model Evaluation and Testing

We evaluate the TIPR model using a variety of metrics. For each binary classifier, the AUROC and AUPRC is calculated. The multi-class MSC classifier is evaluated on sensitivity, specificity, and micro/macro F1 scores, reported in the results section. In addition to standard numerical metrics, we evaluate the model in a more practical setting by predicting TSSs on a larger scale, using entire regions of the genome.

While the model can be successfully applied to predict the probability that an individual nucleotide is a TSS, more commonly we wish to know of *regions* where transcription initiation is likely to occur. To evaluate the TIPR model on a practical

scale, we tested the model on 4 kb regions upstream and downstream of TSSs in the held-out set. First, each nucleotide in the surrounding 8kb region is featurized as above. Next, nucleotides are classified as TSS or Not-TSS using the appropriate TIPR model. The output of this process is a signal reflecting the probability of transcription initiation at every nucleotide scanned. After this signal is produced, it is smoothed using a moving average (10 nt for SP vs BR, 2 nt for SP vs No and BR vs No). As in the single-nucleotide prediction workflow, the SP vs BR model is first used to select the appropriate initiation pattern classifier. Finally, the locations of TSS clusters are determined from the resulting probability signal. The signal is smoothed using a 2 nt moving average, then TSS tag clusters are defined by locating regions of the signal where the probability rises above a probability threshold. A TSS cluster ends after the signal falls below the threshold for 10 consecutive nucleotides.

The above procedure was repeated using a range of probability threshold values between 0.05—0.95. The distance between the predicted TSS and ground-truth TSS locations were calculated, along with the number of correctly predicted TSSs (true positives) and additional positive predictions (false positives). To understand the impact of the flanking region size on the above performance metrics, we also calculated an AUROC-like metric as the flanking distance was increased from 100 bp to 4 kb. This metric is a relative measure of the TPR and FPR of the model, and not directly comparable to standard AUROC values. To calculate these values, a TPR of 1.0 corresponded to all test set TSSs being predicted correctly as above, a FPR of 1.0

corresponded to all flanking nucleotides in the test set being incorrectly predicted as transcribed, and that the MSC model would achieve a TPR and FPR of 1.0 at a probability cutoff of 0.0. The total number of true positives and false positives were computed as described above, and the area under this curve was computed. Note that this is an overestimate of the FPR, as due to the peak-calling procedure, it would be impossible for all nucleotides to be classified individually. However, because the FPR is normalized by the flanking size, these relative AUROC values give insight into the performance of the MSC model as more nucleotides are examined.

3.6 Construction of Synthetic Data Sets

In order to understand the behavior of regularization methods on datasets with multiple optimal feature vector assignments, we created synthetic datasets designed to model expected biological regulatory mechanisms. Our primary purpose of these experiments was to understand what happens when a dataset can be explained by different *equivalence classes* of predictors. For example, a gene could be regulated through the presence of transcriptions TF1 and TF2 *or* by the presence of TF3 and TF4. An ideal regularization method and model would inform us of both these possibilities, instead of simply building a model which utilized TF1 and TF2, while completely ignoring TF3 and TF4 as redundant. The synthetic datasets were designed to be simple, toy-like examples to ease the interpretation of the resulting models, while still modeling most of the complexity of these biological systems. The results of this experiment are reported in section 0.

Three different synthetic datasets were created and analyzed in this study. Seven continuous predictors were used in each dataset, numbered TF1 – TF7 and examples were assigned binary class labels. Each feature was modeled as two Gaussian distributions with one distribution per class label. Feature values were randomly selected from these distributions based on the desired class label. Datasets were balanced with a 50/50 split between negative and positive class labels. In each dataset, two equivalence classes were modeled. TF1 and TF2 formed one group of correlated features, TF3 – TF5 formed another, and TF5 – TF6 were uncorrelated with the class label (or each other). Unique datasets were constructed by generated different combinations of examples and changing the variance and means of individual feature distributions. The parameters used to create each synthetic dataset are listed in Table 2.

Table 2: Parameters used to generate synthetic datasets

The parameters used to generate synthetic data sets containing equivalence classes of predictors. Each predictor has two associated Gaussian distributions. Positive examples are drawn from the distribution defined by the parameters (u_p, σ_p), while negative examples are drawn from (u_n, σ_n). The predictors TF6 and TF7 are not listed in this table, as they were the same across all datasets, and were completely uncorrelated with the class label.

Dataset	**Combinations/Notes** *The groups of TFs from which examples are generated.*	**TF1** **(μ_p, σ_p)** **(μ_n, σ_n)**	**TF2** **(μ_p, σ_p)** **(μ_n, σ_n)**	**TF3** **(μ_p, σ_p)** **(μ_n, σ_n)**	**TF4** **(μ_p, σ_p)** **(μ_n, σ_n)**	**TF5** **(μ_p, σ_p)** **(μ_n, σ_n)**
TF Separate	(TF1-2), (TF3-5) *Features of each example are drawn from either TF1-2 class OR TF3-5 class, but not both.*	(0.8, 0.1) (0.1, 0.01)	(0.8, 0.1) (0.1, 0.01)	(0.6, 0.1) (0.1, 0.01)	(0.6, 0.1) (0.1, 0.01)	(0.6, 0.1) (0.1, 0.01)
TF All Mixed	(TF1-2), (TF3-5), (TF1-2, TF3-5) *Features of each example are drawn from TF1-2 class, TF3-5 class, or both.*	(0.8, 0.1) (0.1, 0.01)	(0.8, 0.1) (0.1, 0.01)	(0.6, 0.1) (0.1, 0.01)	(0.6, 0.1) (0.1, 0.01)	(0.6, 0.1) (0.1, 0.01)
TF All Mixed Diff Variance	(TF1-2), (TF3-5), (TF1-2, TF3-5) *Within an equivalence class, one feature has higher variance than the other.*	(0.8, 0.2) (0.3, 0.01)	(0.8, 0.1) (0.3, 0.01)	(3.6, 0.1) (0.9, 0.01)	(3.6, 0.1) (0.9, 0.01)	(0.3, 0.2) (0.9, 0.01)

Table 2: Parameters used to generate synthetic datasets

3.7 Feature Reduction

In this study, we applied several standard feature reduction techniques to our datasets, including PCA, mutual information, and a novel extension of mutual information called MRMR (Ding and Peng, 2005). A brief review of these techniques is given in Appendix A. Feature reduction was performed on SP and BR datasets containing the flanking features described in section 3.3.1.

4 Results

4.1 TIPR Successfully Predicts Broad Initiation Patterns

Previous high-resolution TSS prediction models focused primarily on the prediction of Narrow Peak TSS initiation patterns (Megraw et al., 2009), reasoning that these promoters are likely more tightly regulated by specific transcription factors, as opposed to non-sequence mechanisms such as histone markers and chromatin structure (Rach et al., 2011). We trained and tested the TIPR model on multiple initiation patterns, using the CAGE mouse dataset (Carninci et al., 2006), filtered as described in the Methods section. The training and testing sets used are shown in Table 3.

Our results show that our model can predict initiation patterns beyond the Narrow Peak class with high accuracy (both high AUROC and AUPRC) from sequence content alone (Table 4, Figure 5, and Figure 6). The models trained on a dataset of an individual initiation pattern perform well, meaning that the model describes a set of TFBS enrichments which well-characterize the initiation pattern used to build and test the model. Logistic regression models (such as the ones used in this study) are composed of a series of weights assigned to each input feature. These weights form the linear combination coefficients used to perform model predictions. The magnitude of these feature weights corresponds to the predictive importance of the corresponding input feature. By examining the feature weights of the models, we see different TFs

weighted more heavily, implying that genes with different initiation patterns are likely regulated by different sets of TFs.

Table 3: CAGE datasets used to train and test TIPR model

Initiation Pattern	Total Tag Clusters	Training Tag Clusters	Testing Tag Clusters
Single Peak	1247 (33%)	998	249
Broad Peak	2497 (66%)	1998	499
All	3744 (100%)	2996	748

Summary of CAGE TSS-Seq dataset used for training and testing the TIPR model after quality filtering was performed.

Table 4: Performance of TIPR's three binary TSS classifiers

Model	*AUROC*	*AUPRC*
SP vs NO	0.99	0.72
BR vs NO	0.99	0.81
SP + BR (ALL) vs NO	0.99	0.82

The results of classifying testing examples with the individual binary classifiers TIPR uses internally. The SP and BR models are trained and tested on a single TSS initiation pattern, while the SP + BR model combines both initiation patterns together, creating a general purpose classifier.

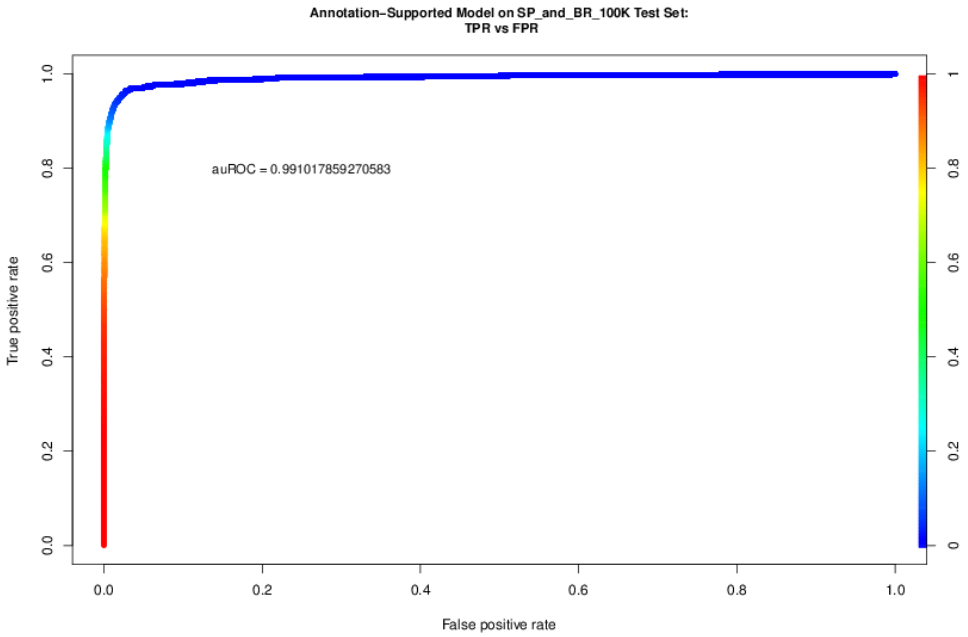


Figure 5: ROC Plot of SP + BR (ALL) vs NO Classifier

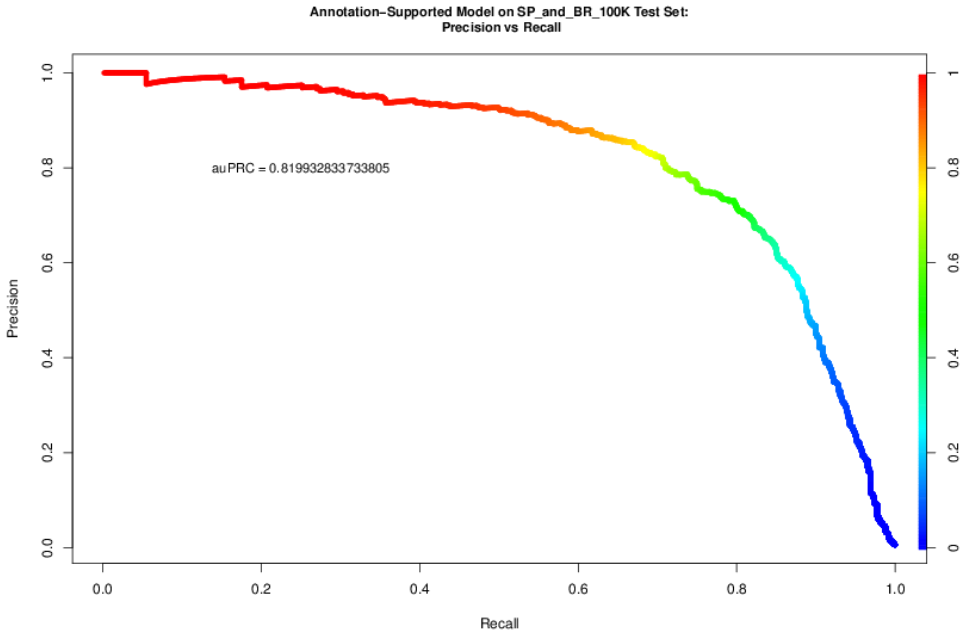


Figure 6: PRC Plot of SP + BR (ALL) vs NO Classifier

While the individual initiation pattern classifiers perform well on their respective datasets, a general purpose classifier is required for the prediction of TSSs without prior knowledge of the initiation pattern. The two general classifiers constructed in this study (ALL and MSC) both performed well at the task of general TSS identification. The ALL model, trained on the combination of both Single Peak and Broad initiation patterns, forms a general-purpose TSS prediction model. Due to this model's simplicity, it is useful for the task of predicting TSSs when the spatial initiation pattern is of little interest to the user. The more complex MSC classifier functions as a general-purpose TSS prediction model, while also providing specific spatial initiation pattern predictions with the same predictive accuracy as the ALL classifier.

As transcription of a gene typically initiates at many locations within a genomic region—as opposed to one single location at a specific nucleotide—a successful model must predict these regions with high resolution and precision. To evaluate the performance of our model in this context, we performed a scanning procedure where the model was used to predict the probability of a TSS at each nucleotide within an 8kb region containing an experimentally observed TSS tag cluster. After smoothing of the probability signal output by the model, we evaluated performance on two metrics: the number of TSSs predicted at a given probability threshold compared to the number of false positives (Figure 7), and the distance between the predicted and ground-truth TSS locations (Figure 8). These results show that the generic ALL models can identify

TSS clusters with high accuracy regardless of initiation pattern. This also demonstrates the necessity of the SP vs NO and BR vs NO classifiers, and how the MSC and ALL classifiers perform better overall than either of these specialized classifiers. The SP classifier curve (diamond, Figure 7) is below all other models, meaning that this model identifies fewer true positives. As SP initiation patterns are less common overall, this is expected, as the SP classifier is trained to identify only Single Peak patterns, while a majority of the dataset is composed of BR TSS tag clusters. On the other hand, Figure 8 shows that the SP classifier is more precise in terms of locating the ground truth TSS. However, as SPs are typically narrower and more well-defined than the BR pattern, this increased resolution may simply be an artifact of the SP TSSs which are correctly identified by the SP model. The Broad initiation pattern classifier identifies TSSs with roughly the same accuracy as the ALL and MSC models, but the preciseness of the BR model is slightly reduced. In this way, the ALL and MSC models perform better than either the SP or BR model alone.

As a numerical comparison, we also compared the MSC and ALL models at specific FPRs (Table 5). At a false positive rate corresponding to 1 false positive per kilobase, the ALL classifier correctly identifies 89.1% (667) of the testing examples, with an average distance of 30 nt between the predicted and actual TSS mode. In comparison, the MSC classifier correctly identifies an additional 11 testing examples, or 90% (678) of the testing examples, while the resolution decreases by 2 nt on average.

To understand how the metrics shown in Figure 7, Figure 8, and Table 5 were affected by the size of the region scanned, we computed a metric we termed relative AUROC, described in detail section 3.5. This metric essentially normalizes the FPR approximated in the x-axis of Figure 7 by the width of the flanking region and computes the area under this curve. As the width of the region scanned surrounding the TSS was increased from 100 bp to 8 kb, the relative AUROC was computed is summarized in Figure 9. The decreased performance with smaller flanking regions (corresponding to a higher false positive rate) is due to the total percentage of the flanking region predicted as being transcribed by the MSC model. Intuitively, regions surrounding the experimentally-supported TSS are more likely to be predicted as transcribed themselves, especially in broad peak initiation patterns. However, our analysis considers all such predictions as false positives, under-estimating the model's sensitivity. When a small region surrounding the TSS, a high percentage of the flanking nucleotides will be predicted as transcribed, increasing the calculated FPR. However, when a larger region is analyzed, a vast majority of the flanking region is (correctly) predicted as non-transcribed, driving down the FPR, and increasing the relative AUROC. Relative AUROC stops increasing after 1000 nt surrounding the TSS are examined. This result implies that the majority of additional predicted TSSs per kilobase fall within 500 nt upstream or downstream of the experimentally supported TSS, with few predictions further upstream or downstream in intergenic and

exonic/intronic regions. A representative example of this scanning procedure is shown in Figure 10 and Figure 11.

Figure 11 shows the output of the model when scanned 4 kb upstream and downstream of the TSS, which shows the model predicts several additional TSSs while labeling most sequence as non-TSS locations. The model predicts a secondary high-probability region slightly further downstream of the putative TSS. While this site is not highly expressed in the CAGE data, there is some evidence of transcriptional activity in addition to several ESTs in the region. A third TSS location is predicted further downstream on an exon boundary and is also supported by several ESTs and a CAGE TSS tag cluster upstream of the predicted TSS. These figures show that TIPR is both accurate and precise, correctly identifying the CAGE-supported TSS tag cluster and predicting very few other TSSs (potential false positives).

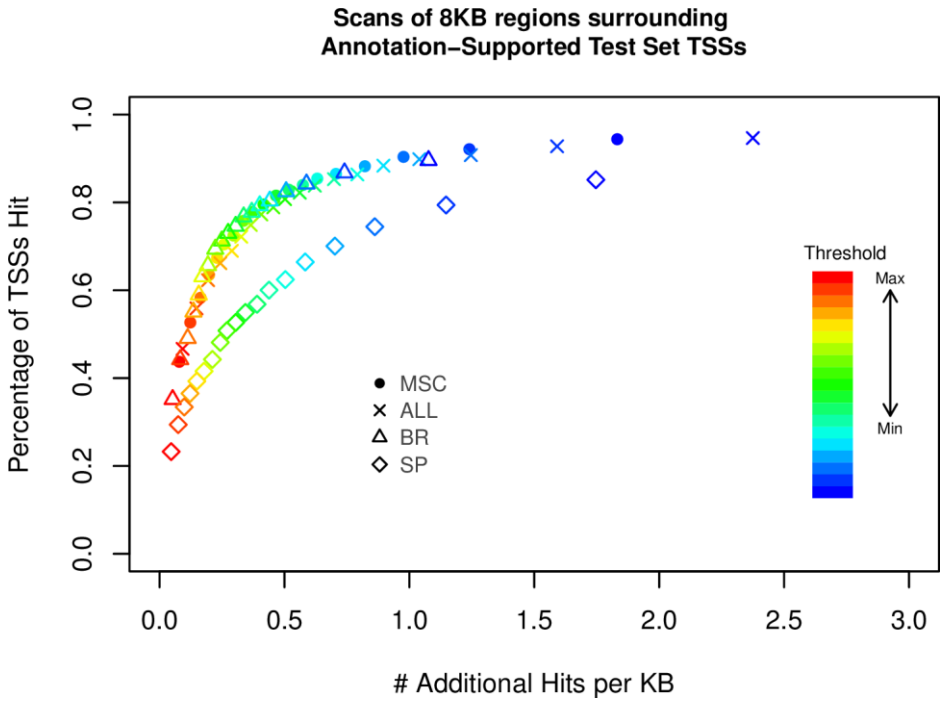


Figure 7: Performance of all classifiers during gene scanning

The accuracy of all classifiers when applied on a large scale to the entire testing dataset. The vertical axis shows the percentage of TSSs in the test set which are correctly predicted (TSSs). The horizontal axis measures the number of additional TSSs which are predicted (false positives). The color scale shows the probability cutoff threshold, the value the model prediction must be above to be considered a TSS.

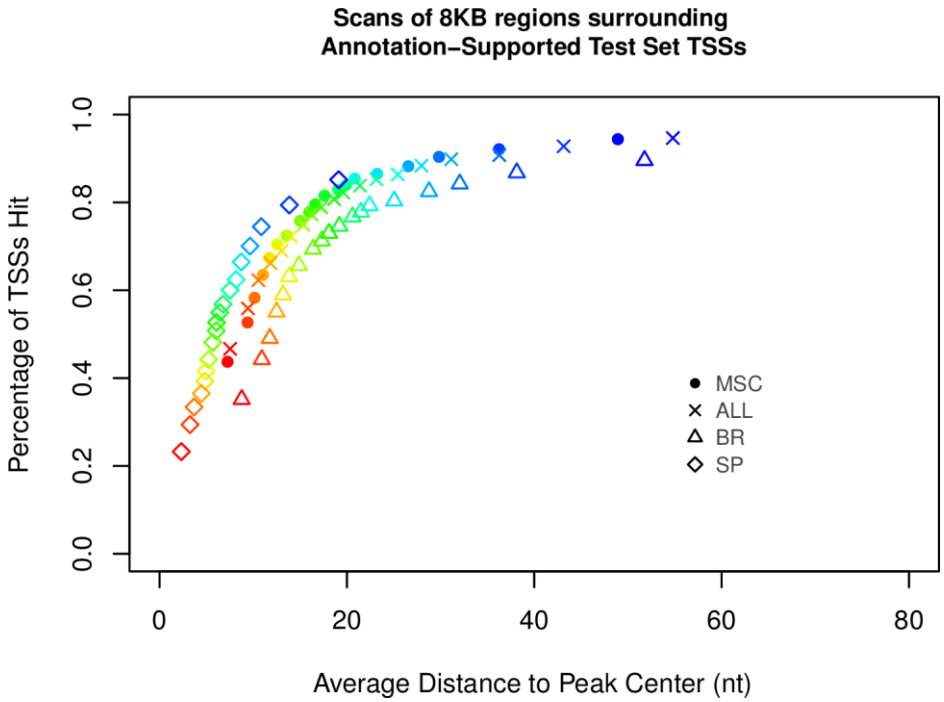


Figure 8: Preciseness of all classifiers during gene scanning

The preciseness of all TSS classifiers, as quantified by the distance between the TSS tag cluster mode (experimentally supported ground truth data) and the center of the predicted tag cluster.

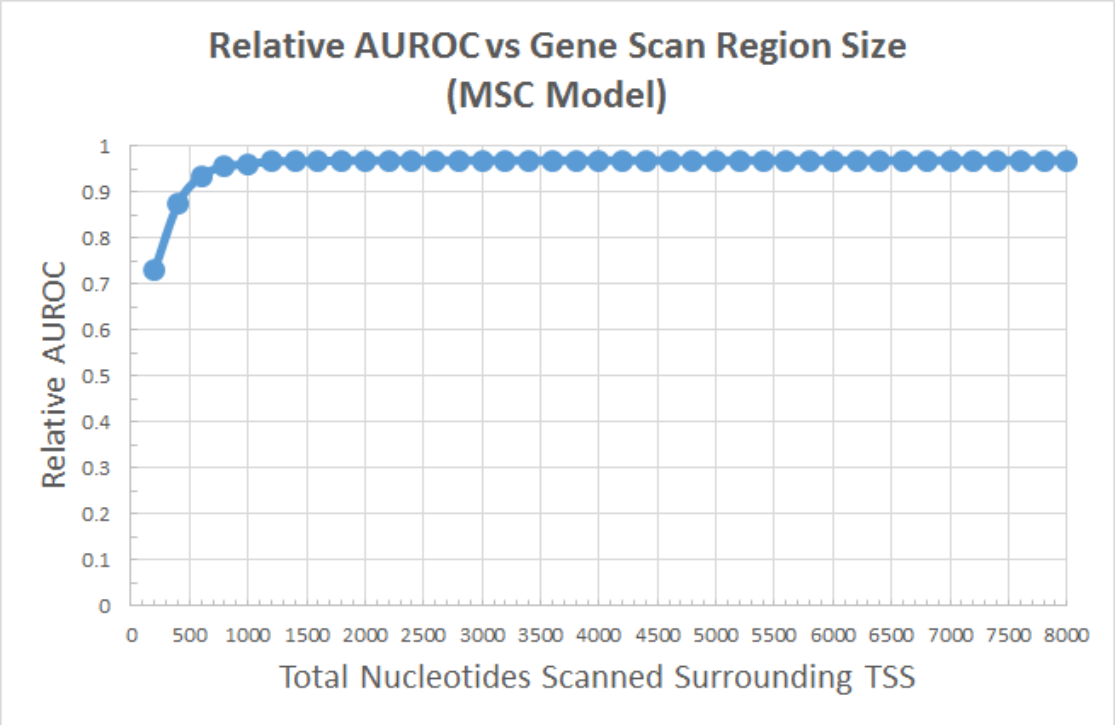


Figure 9: Relative AUROC of MSC model as predicted region is increased

This graph shows how the relative AUROC of the MSC model in gene scanning (see section 3.5: Model Evaluation and Testing for details on calculation), as the region surrounding the experimentally-supported TSS is increased from 200 nt to 8000 nt. This plot shows an increase of relative AUROC as the flanking region is increased, a value which becomes nearly constant when 1000 nt or more (500 nt upstream and 500 downstream) surrounding the TSS are scanned. The lower relative AUROC with smaller flanking regions is due to higher percentages of flanking regions being predicted as transcribed by the MSC model.

Table 5: Comparison of accuracy and resolution of ALL and MSC models

	ALL CLASSIFIER		MULTI-STAGE CLASSIFIER		
ADDITIONAL HITS/KB	Percent TSSs Hit	Distance to Center (nt)	Percent TSSs Hit	Distance to Center (nt)	Additional TSSs Hit
0.76	86.23%	24.61	87.17%	26.09	+7
0.85	87.57%	26.35	88.77%	28.61	+9
1.00	89.17%	30.40	90.64%	32.13	+11
1.25	90.78%	36.22	92.65%	37.51	+14
1.51	91.98%	41.19	93.45%	43.73	+11

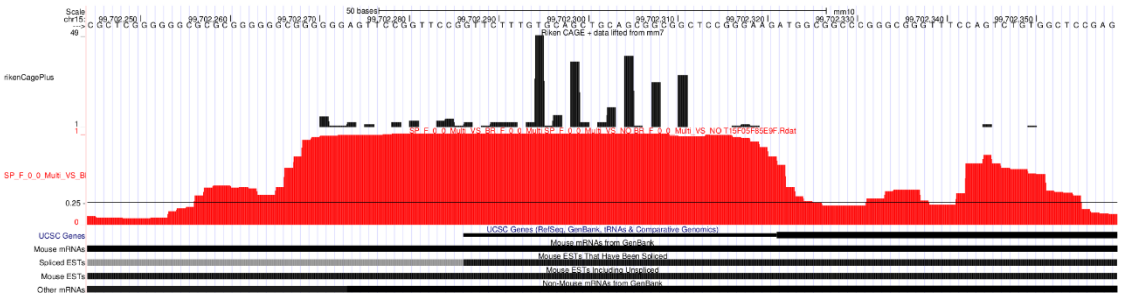


Figure 10: Example of TIPR gene scan output surrounding CAGE TSS

This figure shows the agreement between CAGE TSS-Seq data (top) and TIPR prediction (middle, red) on *M. musculus* gene Smarcd1. The prediction region is centered around the CAGE TSS tag cluster mode and matches the shape of the tag cluster closely.

Figure 11: Example of TIPR gene scan output of 8 kb region of sequence

This figure shows the TIPR MSC model used to predict TSSs in the 8 kb region surrounding the gene Smarcd1. The top track displays the alignment of TSS-Seq (CAGE) reads along chromosome 15 of the *M. musculus* genome (CAGE tag cluster T15F05F85E9F). The second track (in red) is the probability output from the TIPR MSC model. The expanded track below shows that Mouse ESTs align well with TIPR's predictions. Some additional TIPR predictions are located near other CAGE tag clusters or ESTs.

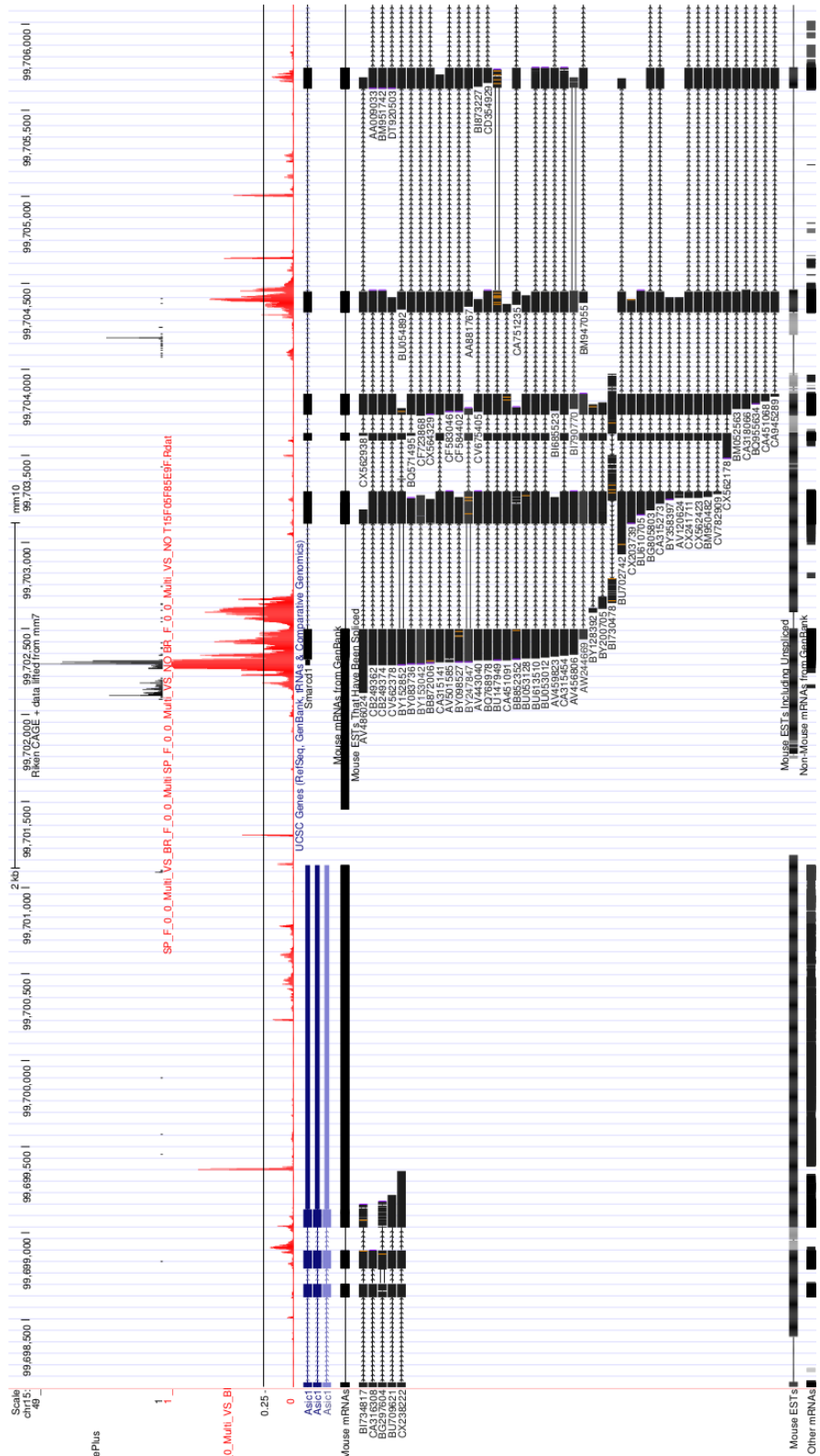


Figure 11: Example of TIPR gene scan output of 8 kb region of sequence

4.2 TIPR Predicts Initiation Pattern Type

In addition to predicting the locations of TSSs from sequence content, our model predicts which initiation pattern the surrounding TSS cluster is likely to form. This is a more complex type of prediction, because the classifier must incorporate information that effectively considers an entire genomic region of nucleotides as possible TSSs. On the held-out test set, the SP vs BR model achieves an average AUROC of 0.88 with an AUPRC of 0.84. This model is combined with the TSS classifiers built on individual initiation patterns to produce the final MSC model. Given a model which can differentiate between initiation patterns, the initiation pattern prediction is used to select the appropriate model to predict if a nucleotide is a TSS. The MSC model achieves the same performance as the binary ALL classifier with regards to precision and recall, if all positive TSS predictions (SP or BR) are considered as true positives. At the cutoff threshold which achieves the optimal F1 score, the MSC model has a recall of 0.84 and precision of 0.84. At the same recall of 0.84, the binary ALL classifier achieves a precision of 0.857. In a multi-class prediction model context, the MSC classifier achieves a macro-F1 score of 0.79. The Macro-F1 statistic is the traditional multi-class F1 score which has been adjusted for the size of each class. Because the negative (No TSS) class contains many more examples than the SP or BR classes, this weighting is important for evaluating classifiers. For example, the unweighted multi-class F1 statistic of the MSC classifier is 0.98, heavily dominated by the size of the No TSS class.

While the MSC classifier performs no better than the simple ALL classifier on the single-nucleotide classification dataset, it does provide important additional information—the predicted spatial initiation pattern of the TSS. This extra information is highly relevant from a biological perspective, as initiation patterns have been shown to associate with different biological interpretations. Genes associated with Single Peak initiation patterns are often tissue-specific developmental genes, while genes with broad patterns are more commonly involved in general and housekeeping processes. The initiation pattern of a gene has been demonstrated to be related to the promoter structure of the gene, including the presence of TFBS elements, sequence enrichments including CpG islands in mammals, and gene function (Carninci et al., 2006; Megraw et al., 2009; Morton et al., 2014; Rach et al., 2009).

By examining the feature weights of these models, we can gain insight into the differences in promoter content of different initiation patterns. Sequence enrichment features were among the most informative features across all the models. The SP vs NO and BR vs NO models highly weighted the GC content feature (representing the presence of CpG islands), while highly negatively weighting the CA enrichment feature, implying that promoter regions may be depleted of CA dinucleotides. While the SP and BR models contain approximately the same number of features (657 for SP vs 609 for BR), the relative importance of TFs varies. For example, the INI motif is the 3rd most important feature for the BR model, while it is ranked as 17th most

important in the SP model. The TATA box motif (highly ranked in SP) does not appear at all in the BR model.

We can also examined the features of the SP vs BR model to better identify differences between the two initiation patterns. This model is simpler (containing only 49 features), but highlights TFs which are likely unique to NP initiation patterns, such as TATA and CDXA, and those which are more associated with broad patterns (such as GABPA and CpG island enrichment). Table 6 shows the magnitude of the top-10 highest-weighted feature coefficients present in the SP vs BR model.

The performance of the predictive models suggests that some of the underlying biological mechanisms which give rise to multiple transcription initiation patterns can be inferred from the models. This interpretability is an important feature if the TIPR model which many other TSS predictive do not provide in as clear a manner. There are two related inquiries which we also investigated. First, we investigated why some testing examples were being incorrectly classified by the model, suggesting these examples used a different set of elements to initiate transcription. Second, we considered ways of improving the interpretability of the model and how different combinations of regulatory elements could impact the resulting model.

Table 6: Top-10 features selected by SP vs BR model

Single Peak		Broad Peak	
Feature	**Weight**	**Feature**	**Weight**
CDXA_02_REV_4	0.221	GABPA_FWD_3	0.131
CDXA_01_REV_4	0.213	GCcontent	0.128
ATATA_B_FWD_4	0.213	E2F_Q2_FWD_6	0.126
TBP_01_FWD_4	0.197	HINFP_REV_3	0.086
CAP_01_FWD_4	0.193	E2F1_Q3_FWD_4	0.069
CDXA_01_FWD_4	0.183	CREB_02_FWD_4	0.060
TBP_Q6_REV_4	0.180	MYB_Q5_01_FWD_7	0.056
GEN_INI_B_FWD_4	0.171	CKROX_Q2_REV_7	0.055
CAP_01_FWD_5	0.138	NRF1_Q6_REV_4	0.044
ATATA_B_FWD_3	0.094	CHCH_01_FWD_5	0.042

These 20 features were identified by the SP vs BR model as the most discriminative features for differentiating SP and BR initiation patterns.

4.3 Analysis of Incorrectly Classified TSSs

While the predictive TSS models performed very well, there were a subset of test examples which could not be successfully predicted. We investigated why these misclassifications occurred to determine the reason for failure and improve the model

if possible. In general, it did not appear that these false negatives were simply caused by the choice of probability threshold. That is, the predicted probability of the misclassified examples of being a TSS was far below the probability cutoff threshold used to differentiate a TSS from a non-transcribed reason, and were instead very close to 0—signifying a confident Non-TSS label (Figure 12). The low-scoring positive examples implied that these promoters were very different from the other promoters in the initiation pattern, more closely resembling the non-transcribed negative sites used in training. To investigate this further, we examined the promoter makeup of these misclassified promoters and compared them to the correctly-classified nominal promoters. Overall, while these promoters were strongly expressed, their promoters did not appear to contain many known binding elements within the expected regions of enrichment.

Figure 12: Density plots of model probability output

Density plots of prediction model output (x axis) in correctly-classified (first column) and incorrectly-classified (second column) test dataset examples. The red and blue plots are the output of the SP and BR models, respectively. When models correctly classify examples (first column), the probability of the class is very close to 1.0 (predicting the example is a TSS). However, when examples are misclassified (second column), both initiation pattern models are very close to 0.0 (predicting the example is not a TSS).

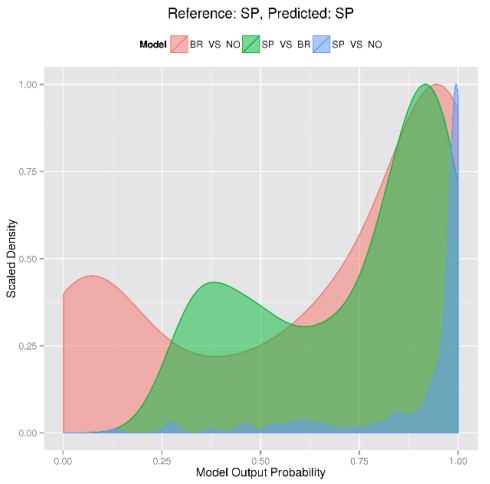
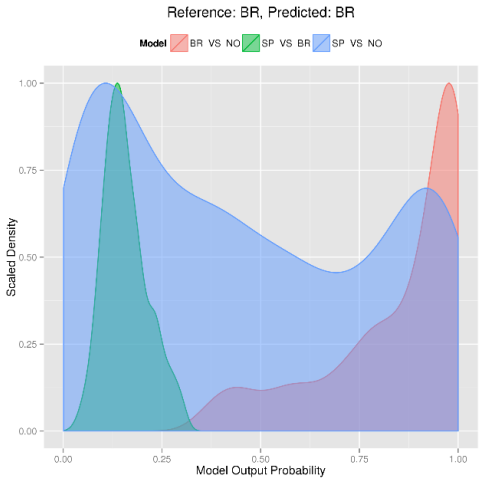
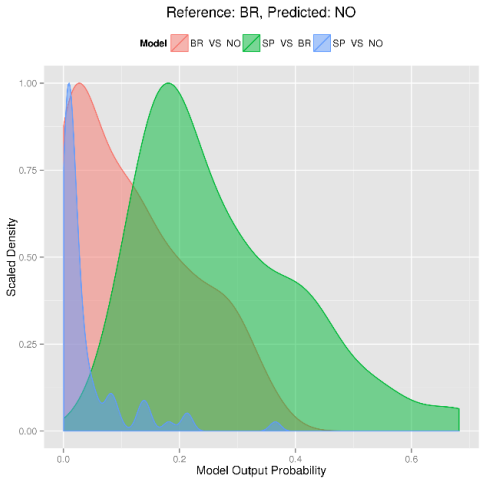
Class	*Correctly Classified*	*Incorrectly Classified*
Single Peak		
Broad Peak		

Figure 12: Density plots of model probability output

We separated promoters into two classes: those which were correctly classified by the model, and those which were incorrectly classified. Within these two classes, we counted the number of TFBSs of each TF located within the regions of enrichment of each example (FPR=0.001), normalized by the size of the ROE and computed the fold decrease in the misclassified set compared to the correctly classified examples. 80/110 of the TFs had a fold decrease greater than 1.0. Included in this list were the highly weighted TFBSs TATA (fold decrease 14.07), ETF (1.39), SP1 (3.15), and KLF1 (3.28). Table 7 lists the fold change and feature weights of the transcription factors with the largest fold decrease. These results indicate that the misclassified examples may have a different promoter structure, as on average they do not contain TFBSs in the same regions as the correctly classified promoters.

Table 7: Changes in promoter composition of TSSs

TFBS	HITS/KB (CORRECTLY-CLASSIFIED)	HITS/KB (MISCLASSIFIED)	FOLD DECREASE	FEATURE WEIGHT IN MODEL
TATA_01	19.93	1.40	14.23	0.05
TBP	19.71	1.40	14.07	0.33
STAT1_03	8.81	1.55	5.69	0.01
HLTF	7.10	1.55	4.59	0.00
EGR1	4.31	0.96	4.47	0.02
FOXD1	5.76	1.40	4.11	0.00
FOXL1	10.36	2.67	3.87	0.00
ZNF263	2.44	0.70	3.48	0.00
ARID3A	5.39	1.55	3.48	0.00

Change in the promoter composition of correctly- and incorrectly- TSSs. A promoter was considered to contain a TFBS (be a “hit”) if a site was located within the TF’s ROE and received a log-likelihood score corrosponding to a FPR of 0.001.

To investigate if these misclassifications were caused by a shift in TFBS locations compared to the majority of TSSs (suggesting a secondary region of functionality), we recomputed the ROEs using only TSSs in the misclassified dataset. If the ROEs

defined by the misclassified dataset were different than those defined by the entire TSS dataset, this would suggest that some TFs have multiple regions relative to the TSS where they are biologically functional. Because the different initiation patterns used different ROEs, we further split the misclassified dataset by their labeled initiation pattern and recomputed ROEs over this new set.

This analysis did not reveal any new potential ROEs for this new dataset. Representative examples typical of the TFBS enrichments are shown in Figure 13.Overall, the discovered ROEs were a subset of those already present in the original set of ROEs. In general, these new ROEs were less well defined, likely because of the small number of misclassified examples and the overall reduction in the number of TFBSs as shown in the above analysis. This does not mean that alternate regions of biological functionality do not exist or are not responsible for the transcription of these genes. However, we were unable to discover them using our modeling approach and database of TFBSs.

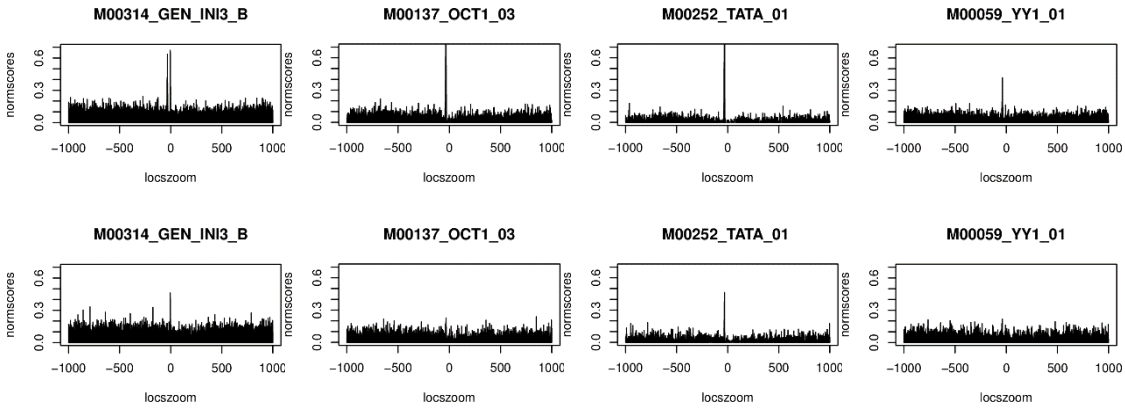


Figure 13: TFBS Enrichment in Correctly and Incorrectly Classified TSSs

This show the changes in the enrichment (y axis) and position (x axis) of TFBS in correctly-classified (top row) and incorrectly classified (bottom row). Overall, misclassified examples had the same locations of maximal enrichment, but these enrichments were less pronounced (Initiator and TATA) or entirely absent (OCT1 and YY1).

4.4 Initiation Patterns Reveal Differences in Gene Promoter Architectures

As previous reported in Megraw et al. (2009), Single Peak initiation patterns have well-defined regions of the promoter with heavy enrichment for specific TFBSs, relative to the rest of the promoter. These enrichments are much less pronounced in the Broad initiation pattern, where only 312/843 TFBSs were detected as containing an ROE on the forward strand, compared to 511/843 in SP. In addition, overall the BR ROEs are not as pronounced or narrowly defined as the ROEs defined by the SP patterns. In most cases, the BR ROEs are a subset of the SP ROEs, though some TFBSs are unique to BR promoters or have different enrichment locations between the

two initiation patterns (Figure 14). In general, the ROEs of broad promoters are wider than single peak promoters, suggesting that while Narrow Peak initiation patterns are likely primarily regulated by the presence of Transcription Factors, and the transcription of Broad patterns is more strongly governed by sequence enrichments.

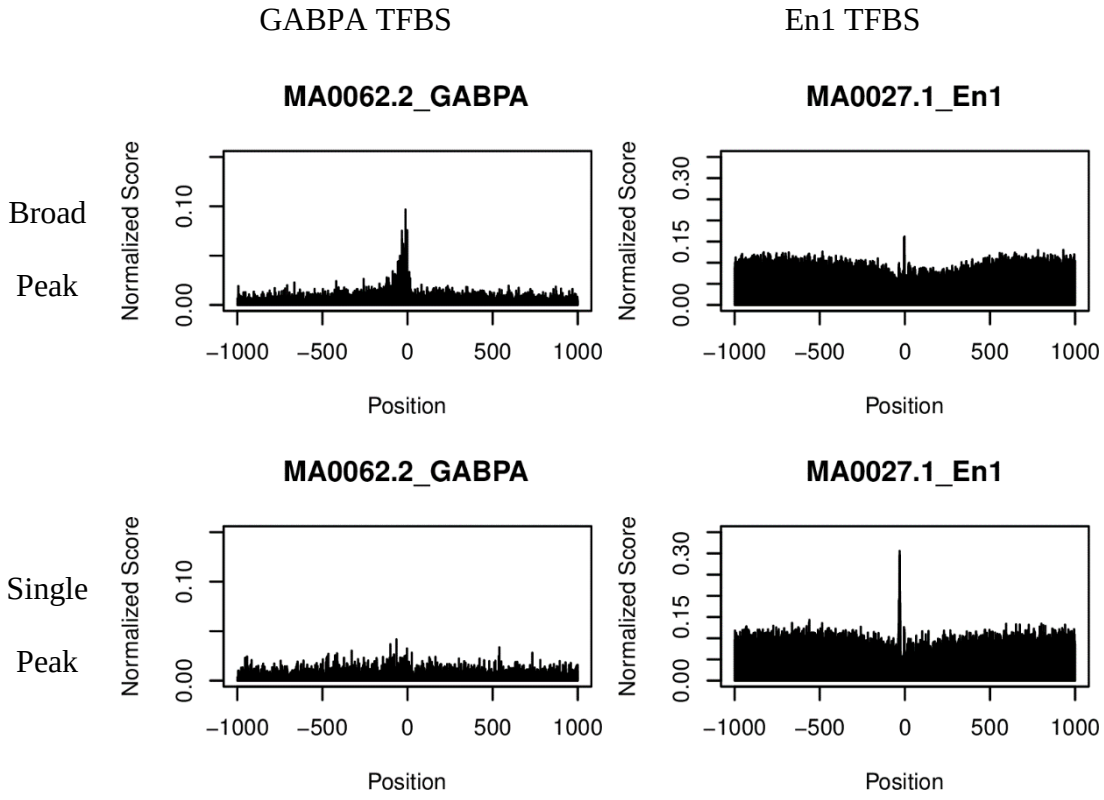


Figure 14: Examples of differentially-enriched TFBSs by initiation patterns

Previous studies have studied the classes of genes associated with different initiation patterns and the differences between these families, including gene function, spatiotemporal expression, and transcriptional regulation (Carninci et al., 2006; Haberle et al., 2014; Morton et al., 2014; Rach et al., 2009). Using a model which

provides the most likely transcription initiation pattern in a region of interest is therefore particularly informative in cases where a gene's functional annotation is incomplete. By making predictions of initiation patterns, we can produce data-informed suggestions of a gene's function or regulatory network. These suggestions can be further improved by combining suggested locations of importance (ROEs) with the feature weights of the TIPR model. We next investigated several techniques to improve the interpretability of the TIPR model.

4.5 Elastic Net Regularization May Improve Model Interpretability

In this work, we evaluated the impact of regularization methods on the features utilized in models to understand how models interpretability could be improved. A more interpretable model can be used to provide more insight into the biological processes underlying transcription initiation.. A brief review of these regularization methods is provided in Appendix B. As described in Methods, synthetic datasets were generated to test the effectiveness of the elastic net in a TSS-prediction context. These datasets were designed to emulate different sets of transcription factors which could regulate a set of genes to understand the behavior of the elastic net in networks of varying complexity. Each synthetic dataset is evaluated on a range of α values ranging from 0 to 1. The final coefficients along with the regularization path are examined. In all cases, models were trained using 5000 instances and tested on a separate held-out test set of another 5000 elements. Because of the simplicity of the synthetic datasets, all tested models (regardless of regularization scheme) achieved 100% accuracy on

their held-out test sets. Therefore all feature weightings assigned by each model are equivalent with regards to classification performance, so maximizing model interpretability is our primary concern in this analysis.

Figure 15 show the clear trade-off between L1 and L2 regularization. This dataset contains examples from two different groups: those which are correlated with features TF1 and TF2, and those which are correlated with features TF3-5. Features TF6 and TF7 are random and uncorrelated. Lasso (α=1) does not include the uncorrelated features in the model, while selecting one feature from each group (TF1 and TF5) as the most important. Note that the other (redundant) features are not completely removed from the model, but are more lowly-weighted with no clear correlation between them and their grouped members along the regularization path. At the other extreme (ridge regularization, α=0), the groupings of the features is obvious from the regularization path, but the uncorrelated features are not completely removed from the model. The elastic net produces almost exactly the same regularization path in a slightly more complex case (Figure 16) where in addition to the groups above, some examples contain features from both the TF1-2 and TF3-5 groups.

We further evaluated regularization methods with more complex, real-world-like datasets. The TF_ALL_MIXED_HIGH_VAR_DIFF dataset tested the ability of regularization methods to avoid higher-variance features when a correlated higher-quality feature was available. In this dataset, TF1 and TF5 (still grouped as described above) had higher variance than the other features within their groups. The

regularization paths of models trained on this dataset are shown in Figure 17. L1 (α=1) was effective at rejecting the nosier variables, assigning them lower weights than their lower-variance group members. TF2 is assigned 6x the weight of the noisy feature TF1, while TF3 and TF4 were assigned weights 12x and 3x larger respectively than the noise predictor TF5. However, this is not the case with ridge regression (α=0), where TF3-5 were assigned nearly identical weights (standard deviation of 0.0032). TF2 was weighted higher than TF1, by only by a factor of 1.3. At larger values of α, the elastic net did weight TF2 over TF1 more clearly while still dropping the uncorrelated features TF6-7. It however did not weight TF5 significantly lower than its lower-variance group members.

These results suggest that while elastic net regularization provides some benefit over L1 and likely can be used to create models which provide insights into equivalence classes of features, L1's sparse model and ability to reject noisy variables provide benefits as well. The elastic net (and L2) perform well at grouping correlated features along the regularization path in simple cases, however it's not clear if these simplistic networks appear in biological networks, or how these regularization methods function with more complex and real-world networks. Even with these limitations, other features of the elastic net—such as the ability to select more predictors than examples—are beneficial, particularly in models with large feature sets where classical feature reduction techniques are impractical. As an alternative to the elastic net, we discuss other feature reduction techniques in the next section.

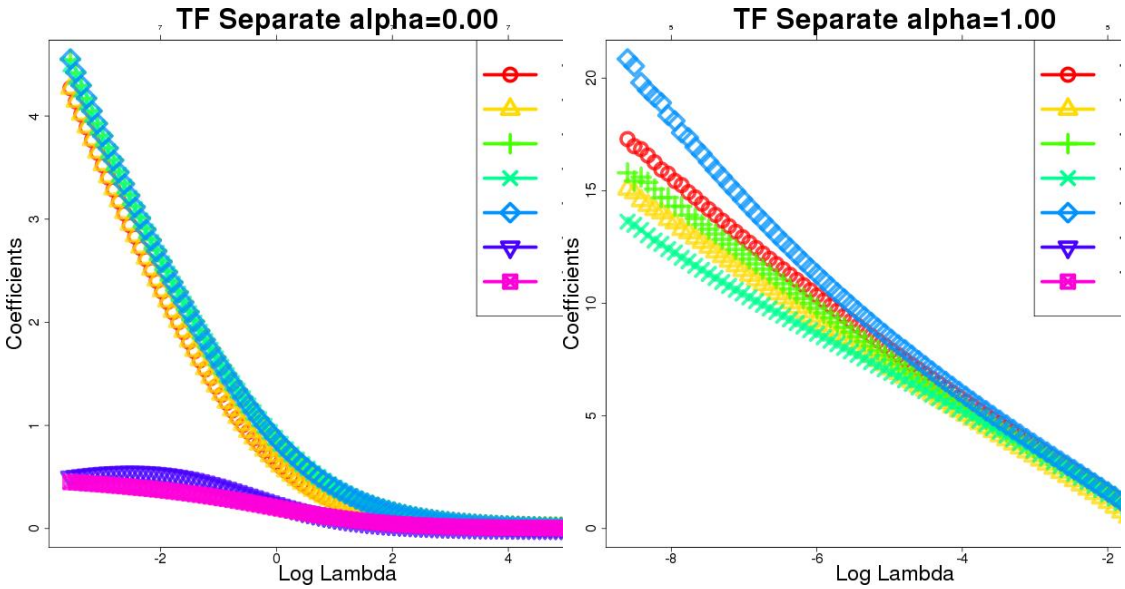


Figure 15: L1 and L2 regularization paths on synthetic TF Separate dataset

The differences between the regularization paths of L2- (left) and L1-regularized (right) regression. In this dataset, the features TF1 and TF2 formed one set of correlated predictors, while TF3-5 formed another. Positive examples contain features from one group class or the other, but not both. L1 (lasso) top weights to 1 feature from each group (blue and red), while the other correlated features receive lower weights, with no clear grouping. Ridge regression shows 2 distinct regularization paths for the 2 feature groups, with each feature within a group being assigned a nearly-equal weight.

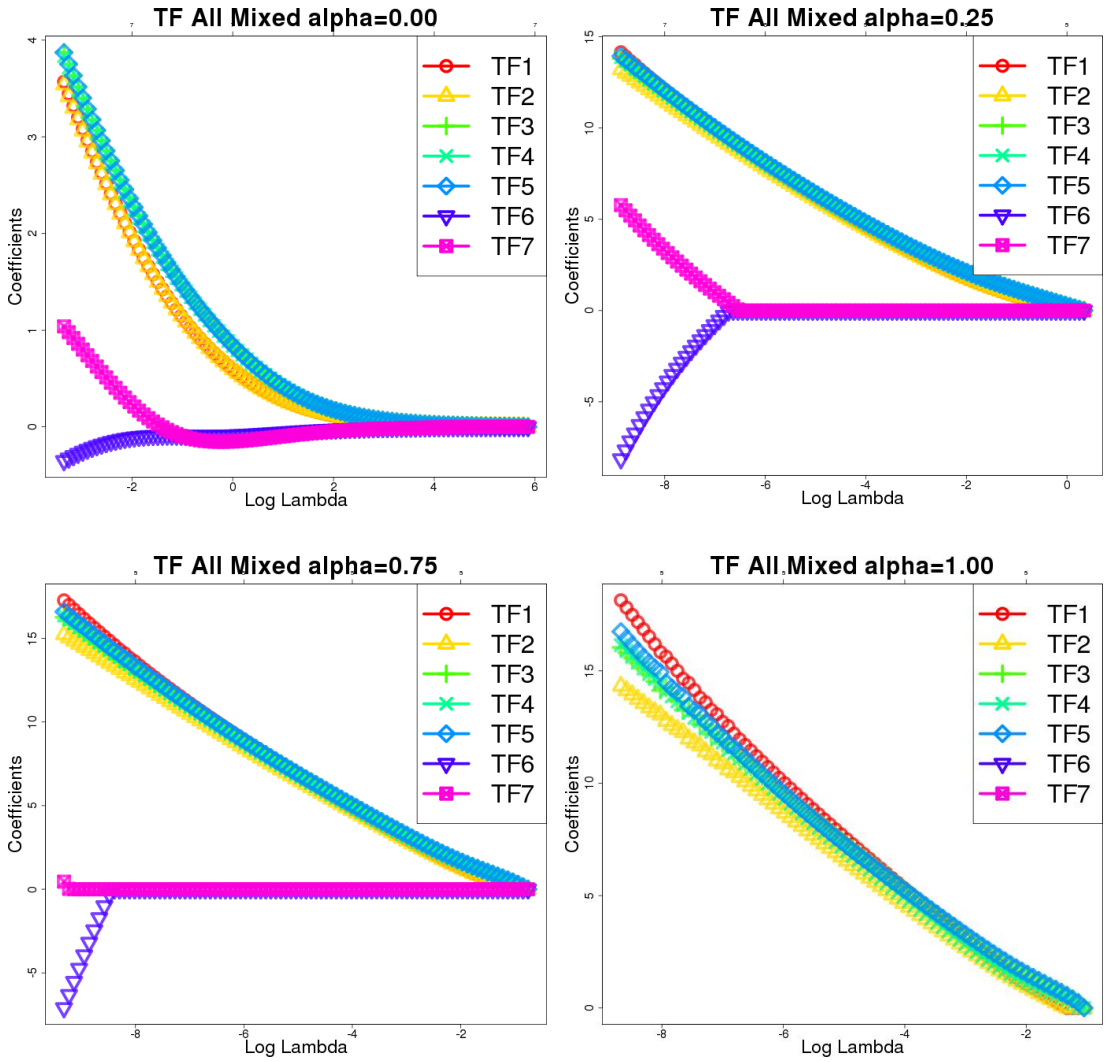


Figure 16: Regularization paths on synthetic TF All Mixed dataset

Regularization paths of ridge (top-left), lasso (bottom-right), and elastic net regularized regression on a dataset containing positive examples with features from each feature group, or both groups at the same time. As in Figure 15, L1 selects 1 feature from each group as the most important (TF1 and TF5). When the L1- and L2-penalties are combined with the elastic net, the regularization paths of correlated features grow together as the L2- penalty is increased (α becomes smaller). However, the uncorrelated features (TF6-7) appear more highly weighted in this model than in the one shown in Figure 15 as α is decreased.

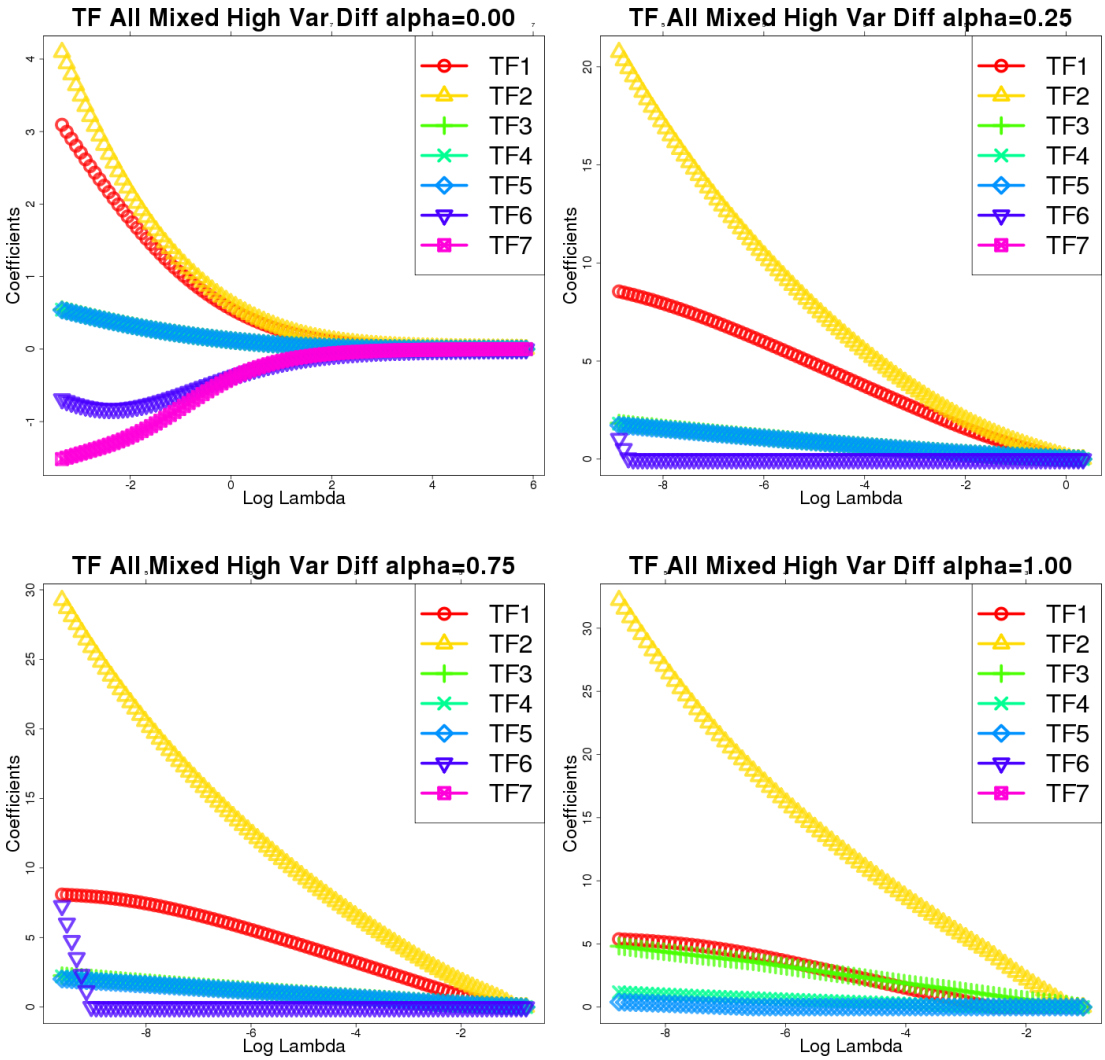


Figure 17: Regularization paths of synthetic TF All Mixed High Variance dataset

The regularization paths of the elastic net in a dataset containing correlated, but high-variance (noisy) features. L1 is successful at rejecting the high-variance features TF1 (red) and TF5 (light blue), favoring their lower-variance group members instead by assigning them higher weights.

4.6 Feature Reduction

Due to the large number of features included in our model and L1's issues with large numbers of predictors, we investigated feature reduction as a way to reduce the number of features considered by the model. While L1 can be an effective feature selection technique, building a regression model with large numbers of features is likely not as efficient as performing some pre-processing before model construction. This step was especially important when building models which contained all pairwise combinations of features, as these models contained millions of features. We evaluated PCA and Mutual Information-based techniques in this study. Initial testing revealed that it was impractical to apply PCA to our datasets directly due to the number of features (over 50 millions).

Our primary question was if pairwise features contributed extra information which would improve the performance on misclassified examples. Because information gain does not consider interactions between features, it can be computed relatively efficiently and in parallel (unlike PCA). We began by computing the information gain of each feature (including pairwise combinations) in the SP vs No TSS dataset. In addition, these datasets included flanking features covering 50 nt upstream and downstream of the TSS, with 1 flanking feature set located every 10 nt. Table 8shows the top 6 features, ranked by their information gain. It's obvious from this table that the redundancy introduced by flanking and pairwise features limits the usefulness of this method for feature reduction. From this analysis, it's clear that ETF, E2F, and GC

sequence enrichment are all informative features. For example, GC content is a good predictor of TSSs. At the same time, the pairwise (GC content)x(GC content) feature is equally informative. In addition, sequence enrichment flanking features are all highly correlated as they are computed over a wide window (250 nt), so the pairwise interactions between sequence enrichment features are redundant. While this situation could be avoided for sequence enrichment features by not computing flanking features or pairwise interactions for them, these same problems occur with TFBS features as well. For example, the table shows that E2F is an informative feature with high information gain. This causes pairwise combinations of features with E2F to be biased by E2F's information gain. Unfortunately, this redundancy limits the use of conventional mutual information for feature reduction.

Table 8: Information Gain of pairwise and flanking features

Feature	Info. Gain
M00695_ETF_Q6_FWD_3_2_x_M00803_E2F_Q2_REV_4_-4	0.375
M00695_ETF_Q6_FWD_3_2_x_M00803_E2F_Q2_REV_4_-3	0.370
M00695_ETF_Q6_FWD_3_2_x_M00803_E2F_Q2_REV_4_-2	0.370
M00695_ETF_Q6_FWD_3_1_x_M00803_E2F_Q2_REV_4_-4	0.367
M00695_ETF_Q6_FWD_3_2_x_M00803_E2F_Q2_REV_4_-1	0.366
GCcontent_-8_x_GCcontent_5	0.362

Several methods have been proposed for feature reduction in datasets with highly correlated features (Ding and Peng, 2005; Hall, 1999; Liu and Motoda, 2007; Vinh et al., 2012; Xing et al., 2001; Yu, 2004; Yu and Liu, 2004). In summary, these methods build on conventional mutual information by ranking features by their information gain, and then calculate the pairwise mutual information between all features, selecting the best non-redundant feature at each step. We used the technique proposed by Ding and Peng (2005), a method called MRMR (minimum redundancy, maximum relevance). The algorithm is briefly described in Appendix A.

This procedure avoids the issues with mutual information described above with redundant features by selecting the least-redundant remaining feature. Unfortunately it requires the calculation of mutual information between each unselected predictor and

all of those which have already been selected. In addition, MRMR cannot be parallelized because of the sequential nature of the algorithm. We first tested the MRMR algorithm on the SP vs No TSS classification model, containing 150,000 features. After approximately 120 hours, the algorithm had selected only 760 features (2.2 GHz CPU with 512 GB RAM). We evaluated these selected features by building an SP vs No TSS classifier using only these features and compared the performance to the original model containing 150,000 features. The original model out-performed the feature-reduced model in both AUROC and AUPRC, achieving 0.99 and 0.92, respectively. The feature-reduced model had an AUROC of 0.97 and AUPRC of 0.87. These results show that MRMR is effective at selecting informative features, creating a model using 0.05% of the full models features with only a 5% reduction in AUPRC. However, the speed of the algorithm limits its usefulness on datasets with large numbers of features. For this reason, it was impractical to repeat this experiment using the full set of pairwise-combinations.

Feature reduction techniques are important to improving model performance and interpretability, especially in large datasets where the types of information and the optimal representation of that information is not clear. These results show that MRMR and related techniques show promise, but are in general too slow for large datasets. Other, more efficient techniques will be required for datasets containing tens or hundreds of thousands of features. Regression regularization techniques which

perform variable selection such as the elastic net are an alternative method, and may be more successful than mutual-information based approaches.

5 Discussion

Transcription Start Site prediction has many practical applications, particularly in organisms with poorly annotated genomes. Predictions can be used to assist in the identification of the regulatory networks controlling genes by identifying which TFBSs are positioned in biologically relevant locations relative to the predicted TSS. These models can also be used to identify potential alternative start sites and the regulators which may control these different sites, leading to the production of different isoforms. Many genes have been shown to have tissue-specific transcription start sites (Fürbass et al., 1997; Shemer et al., 1992; Toffolo et al., 2007; White et al., 1998), and different regulatory networks of transcription factors have been implicated in at least some of these genes (Toffolo et al., 2007; White et al., 1998). Another recent study showed a change in TSS selection, initiation pattern, and TF usage during the transition from maternal to zygotic transcription in zebra fish (Haberle et al., 2014). TSS prediction tools can be used to identify potential alternative TSSs, which can help guide wet-lab experiments to validate sites and regulatory networks. The prediction of spatial TSS initiation pattern along the genome can also provide insight into the nature of transcripts produced from the site. For example, it may suggest

spatiotemporal expression more consistent with housekeeping functions or one more consistent with tissue or time-specific expression.

5.1 Improved Model Performance Enhances Genome-Wide TSS Identification

The TIPR model provides a large boost in performance over previous sequence-based models (Megraw et al., 2009). Likely, this is due to an increased number of TFBS PWMs (the complete TRANSFAC dataset) along with new sequence enrichment features. For example, the CA sequence enrichment feature (not included in the S-Peaker model) was highly negatively weighted, implying that promoter regions may be significantly depleted of CA. In humans, CA is known to be the most common simple-sequence repeat motif, with 19.4 repeats occurring per Mb (Hui et al., 2005). Several studies have shown that intronic CA repeats play a role in the regulation of alternative splicing in some genes (Hui et al., 2005; Yang et al., 2013). Sawaya et al. (2013) report that the AC motif is significantly depleted directly downstream of human TSSs, but the same depletion is not seen in the entire promoter region.

This increased performance is crucial for good performance in genome-scale TSS prediction. A successful TSS predictor must both be sensitive and specific, as it is important to predict both transcribed and non-transcribed sites accurately. A model which predicts TSS regions with high sensitivity will correctly identify regions which contain TSSs. However, without high specificity, the resolution of these predictions

will be limited, especially if a large (non-transcribed) region surrounding the true TSS cluster is also incorrectly predicted as a TSS. Our model is capable of identifying TSS regions both accurately and precisely, regardless of the initiation pattern, an improvement over previous models which were focused on a single initiation type or lacked the ability to identify TSSs with high precision.

The ability to predict spatial TSS initiation patterns is a new and novel ability of the TIPR model, something not provided by previous models. In addition to suggestion gene function and potential spatiotemporal expression of specific TSSs, this provides insights into the biological differences between Single Peak and Broad Peak patterns, and what causes them to arise. This information can guide the development of more accurate and informative TSS prediction models, such as high resolution models predicting the level of transcription of genes at the nucleotide level. Differentiating initiation patterns requires a set of features which capture the underlying biological processes which cause these patterns to arrive, along with proper techniques to tune models to identify all patterns successfully. Our works suggests that in addition to TATA, the presence of other TFs such as CDXA and CAP appear to be indicative of narrow peak initiation patterns as well. Future work could build on this model by investigating other feature engineering methods to capture additional sequence information, such as the spatial positioning of TFBSs relative to one another and the region under investigation. Such additions could further boost the performance of this

classifier and provide more insight into the biological rules which cause these initiation patterns to arise.

5.2 Regularization Techniques Can Improve Model Interpretability

Our investigation of different regularization techniques with the goal of improving model interpretability suggest that the elastic net has potential over both ridge and lasso regression. However, the use of the elastic net introduces several new issues which must be addressed. The method requires an additional parameter α, controlling the combination of L1 and L2 penalties. This can be chosen through cross-validation along with the penalty parameter λ at the cost of increased training time. While α will certainly have an impact of model performance, it's important to remember that we wish to optimize for model interpretability in addition to performance. Unlike classification performance, which can be characterized by any number of well-understood statistical metrics, model interpretability is a much more subjective. While synthetic datasets like the ones used in this work can be created to develop an intuition of the elastic net's behavior with correlated features, this may not transfer to real-world datasets where the correlations are not known a priori. Existing datasets with known correlations (derived from wet-lab experiments, or those manipulated to contain correlations, for example) could be used to guide the selection of the α parameter to increase model interpretability.

In addition to the problem of parameter selection, techniques for extracting interpretable information from the model must be improved as well. Our results show that correlated features often follow similar regularization paths, in agreement with Zou and Hastie (2005). One potential method to cluster features by the similarity in their regularization paths, measured by the values of their coefficients at similar λ values. These clustering could form potential equivalence classes which could then be further investigated by retraining models which include only subsets of the classes and measuring the impact on model performance.

6 Conclusion

In this work, we have proposed a new machine-learning based TSS prediction model, capable of identifying TSSs with high accuracy and resolution, along with the predicted spatial initiation pattern the TSS will form along the genome. We have shown that it is possible to predict TSSs of different initiation patterns (including broad peaks) from sequence content alone. In addition to performing well, the TIPR model is easy to interpret. This modeling technique can be used to get new insights into the structure of promoters, the regulation of genes, and what differentiates the genes utilizing different initiation patterns. The TIPR model and techniques therein have applications both within and outside of the field of TSS prediction. Accurate TSS predictions made using sequence content alone can be used to improve genome annotations, particularly in organisms which are poorly studied or annotated. These prediction can be used to suggest TFs which are involved in the regulation of a gene by identifying TFBSs which are correctly positioned relative to the TSS. They can also identify alternative TSSs which may yield different protein products, such as an alternate TSS which skips the first exon of a gene. In summary, these techniques can be used to build general purpose TSS predictors which function across a wide array of species, and have the potential to help open new avenues of discovery in the field of regulatory genetics.

We also investigated feature reduction techniques, the incorporation of new features and datasets, and how different regularization techniques can be used to build

informative models. Many new large scale datasets are regularly being published and have the potential to further improve TSS prediction and to enhance our understanding of transcription regulation (Neph et al., 2012; Thurman et al., 2012). Including these new sources of data into prediction models will require new feature modeling approaches to incorporate data in a manner which is informative to the model. In this work, we investigated multiple feature engineering approaches, including pairwise combinations of features, inclusion of information from regions flanking the location of interest, and methods of incorporation conservation information into the model.

As the amount of available data grows, feature reduction will become increasingly important. Not only is feature reduction important for improving model performance and interpretability, but it is also as a matter of practicality. While models containing thousands or tens of thousands of features can be trained relatively quickly (in hours using currently available computing power), a model containing several hundred thousand features can takes weeks to train. Increased features also increase dataset size, imposing other practical limitations such as the amount of data which can be stored and analyzed at one time. We investigated several feature reduction techniques, with special focus on methods applicable to large datasets containing highly correlated features. MRMR and related techniques showed promise, but the time complexity of the algorithm was too large to be directly applicable. New feature reduction methods will be required as the amount of available data grows.

Regularization methods also warrant more investigation, as it has a direct impact on both classification accuracy and model interpretability. In most modeling techniques, there is a trade-off between accuracy and interpretability. For example, while SVMs, multi-layer neural networks, and decision trees may all capture complex decision boundaries, inferring the underlying biological functions from these models is often difficult. On the other hand, a logistic regression model is simple to interpret when appropriately regularized, but the flexibility of the decision boundary is limited. To increase the usefulness of predictive models for biological inference, we need models which can identify the subgroups of features involved in a process, not just a large collection of features with no apparent sub-structure. The elastic net seems to show some promise in this area with its ability to groups of correlated features. However, more work is needed to develop methods and heuristics to identify these groups to create truly interpretable models.

Finally, there are many other interesting avenues of investigation beyond TSS prediction to which these methods can be applied. The prediction of gene expression (or simply the level of mRNA production) from sequence content has been a challenging problem, with little progress made in recent years (Beer and Tavazoie, 2004; Yuan et al., 2007). The move from a problem of classification to one of regression introduces a host of new challenges, including featurization, model evaluation, and even how training examples should be aligned with each other (as there is no obvious “start” position as there is with the TSS mode used in TIPR).

However, the ability to accurately predict expression from content—and to understand the underlying regulatory mechanisms—would be incredibly helpful for understanding regulatory networks and guiding wet-lab experiments.

Bibliography

Abeel, T., Peer, Y.V. de, Saeys, Y., 2009. Toward a gold standard for promoter prediction evaluation. Bioinformatics 25, i313–i320. doi:10.1093/bioinformatics/btp191

Alam, T., Medvedeva, Y.A., Jia, H., Brown, J.B., Lipovich, L., Bajic, V.B., 2014. Promoter Analysis Reveals Globally Differential Regulation of Human Long Non-Coding RNA and Protein-Coding Genes. PLoS ONE 9, e109443. doi:10.1371/journal.pone.0109443

Beer, M.A., Tavazoie, S., 2004. Predicting gene expression from sequence. Cell 117, 185–198.

Boer, C.G. de, Bakel, H. van, Tsui, K., Li, J., Morris, Q.D., Nislow, C., Greenblatt, J.F., Hughes, T.R., 2014. A unified model for yeast transcript definition. Genome Res. 24, 154–166. doi:10.1101/gr.164327.113

Bucher, P., 1990. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J. Mol. Biol. 212, 563–578. doi:10.1016/0022-2836(90)90223-9

Bucher, P., Trifonov, E.N., 1986. Compilation and analysis of eukaryotic POL II promoter sequences. Nucleic Acids Res. 14, 10009–10026.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V.B., Brenner, S.E., Batalov, S., Forrest, A.R.R., Zavolan, M., Davis, M.J., Wilming, L.G., Aidinis, V., Allen, J.E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R.N., Bailey, T.L., Bansal, M., Baxter, L., Beisel, K.W., Bersano, T., Bono, H., Chalk, A.M., Chiu, K.P., Choudhary, V., Christoffels, A., Clutterbuck, D.R., Crowe, M.L., Dalla, E., Dalrymple, B.P., de Bono, B., Della Gatta, G., di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C.F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T.R., Gojobori, T., Green, R.E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T.K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S.P.T., Kruger, A., Kummerfeld, S.K., Kurochkin, I.V., Lareau, L.F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Madan Babu, M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K.C., Pavan, W.J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J.F., Ring, B.Z.,

Ringwald, M., Rost, B., Ruan, Y., Salzberg, S.L., Sandelin, A., Schneider, C., Schönbach, C., Sekiguchi, K., Semple, C. a. M., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S.L., Tang, S., Taylor, M.S., Tegner, J., Teichmann, S.A., Ueda, H.R., van Nimwegen, E., Verardo, R., Wei, C.L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S.M., Teasdale, R.D., Liu, E.T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J.S., Hume, D.A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., Hayashizaki, Y., FANTOM Consortium, RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group), 2005. The transcriptional landscape of the mammalian genome. Science 309, 1559–1563. doi:10.1126/science.1112014

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A.M., Taylor, M.S., Engström, P.G., Frith, M.C., Forrest, A.R.R., Alkema, W.B., Tan, S.L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S.M., Wells, C.A., Orlando, V., Wahlestedt, C., Liu, E.T., Harbers, M., Kawai, J., Bajic, V.B., Hume, D.A., Hayashizaki, Y., 2006. Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet. 38, 626–635. doi:10.1038/ng1789

Ding, C., Peng, H., 2005. Minimum redundancy feature selection from microarray gene expression data. J. Bioinform. Comput. Biol. 3, 185–205.

Fickett, J.W., Hatzigeorgiou, A.G., 1997. Eukaryotic Promoter Recognition. Genome Res. 7, 861–878. doi:10.1101/gr.7.9.861

Fürbass, R., Kalbe, C., Vanselow, J., 1997. Tissue-specific expression of the bovine aromatase-encoding gene uses multiple transcriptional start sites and alternative first exons. Endocrinology 138, 2813–2819. doi:10.1210/endo.138.7.5257

Haberle, V., Li, N., Hadzhiev, Y., Plessy, C., Previti, C., Nepal, C., Gehrig, J., Dong, X., Akalin, A., Suzuki, A.M., van IJcken, W.F.J., Armant, O., Ferg, M., Strähle, U., Carninci, P., Müller, F., Lenhard, B., 2014. Two independent

transcription initiation codes overlap on vertebrate core promoters. Nature advance online publication. doi:10.1038/nature12974

Hall, M.A., 1999. Correlation-based feature selection for machine learning. The University of Waikato.

Hoerl, A., Kennard, R., 1988. Ridge Regression, in: Encyclopedia of Statistical Sciences. Wiley, New York, pp. 129–136.

Hui, J., Hung, L.-H., Heiner, M., Schreiner, S., Neumuller, N., Reither, G., Haas, S.A., Bindereif, A., 2005. Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. EMBO J. 24, 1988–1998. doi:10.1038/sj.emboj.7600677

Jin, V.X., Singer, G.A.C., Agosto-Pérez, F.J., Liyanarachchi, S., Davuluri, R.V., 2006. Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs. BMC Bioinformatics 7, 114. doi:10.1186/1471-2105-7-114

Knudsen, S., 1999. Promoter2.0: for the recognition of PolII promoter sequences. Bioinformatics 15, 356–361. doi:10.1093/bioinformatics/15.5.356

Koh, K., Kim, S.-J., Boyd, S., 2007. An Interior-Point Method for Large-Scale L1-Regularized Logistic Regression. J Mach Learn Res 8, 1519–1555.

Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H., Bourque, G., 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat. Genet. 42, 631–634. doi:10.1038/ng.600

Leslie, C., Eskin, E., Noble, W.S., 2002. The spectrum kernel: a string kernel for SVM protein classification. Pac. Symp. Biocomput. Pac. Symp. Biocomput. 564–575.

Liu, H., Motoda, H., 2007. Computational Methods of Feature Selection. CRC Press.

Megraw, M., Pereira, F., Jensen, S.T., Ohler, U., Hatzigeorgiou, A.G., 2009. A transcription factor affinity-based code for mammalian transcription initiation. Genome Res. 19, 644–656. doi:10.1101/gr.085449.108

Morton, T., Petricka, J., Corcoran, D.L., Li, S., Winter, C.M., Carda, A., Benfey, P.N., Ohler, U., Megraw, M., 2014. Paired-End Analysis of Transcription Start Sites in Arabidopsis Reveals Plant-Specific Promoter Signatures. Plant Cell Online tpc.114.125617. doi:10.1105/tpc.114.125617

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., Maurano, M.T., Humbert, R., Rynes, E., Wang, H., Vong, S., Lee, K., Bates, D., Diegel, M.,

Roach, V., Dunn, D., Neri, J., Schafer, A., Hansen, R.S., Kutyavin, T., Giste, E., Weaver, M., Canfield, T., Sabo, P., Zhang, M., Balasundaram, G., Byron, R., MacCoss, M.J., Akey, J.M., Bender, M.A., Groudine, M., Kaul, R., Stamatoyannopoulos, J.A., 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature 489, 83–90. doi:10.1038/nature11212

Ni, T., Corcoran, D.L., Rach, E.A., Song, S., Spana, E.P., Gao, Y., Ohler, U., Zhu, J., 2010. A paired-end sequencing strategy to map the complex landscape of transcription initiation. Nat. Methods 7, 521–527. doi:10.1038/nmeth.1464

Ohler, U., Stemmer, G., Harbeck, S., Niemann, H., 2000. Stochastic segment models of eukaryotic promoter regions. Pac. Symp. Biocomput. Pac. Symp. Biocomput. 380–391.

Ohler, U., Wassarman, D.A., 2010. Promoting developmental transcription. Development 137, 15–26. doi:10.1242/dev.035493

Prestridge, D.S., 1995. Predicting Pol II promoter sequences using transcription factor binding sites. J. Mol. Biol. 249, 923–932. doi:10.1006/jmbi.1995.0349

Rach, E.A., Winter, D.R., Benjamin, A.M., Corcoran, D.L., Ni, T., Zhu, J., Ohler, U., 2011. Transcription Initiation Patterns Indicate Divergent Strategies for Gene Regulation at the Chromatin Level. PLoS Genet 7, e1001274. doi:10.1371/journal.pgen.1001274

Rach, E.A., Yuan, H.-Y., Majoros, W.H., Tomancak, P., Ohler, U., 2009. Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. Genome Biol. 10, R73. doi:10.1186/gb-2009-10-7-r73

Rätsch, G., Sonnenburg, S., Schölkopf, B., 2005. RASE: recognition of alternatively spliced exons in C.elegans. Bioinformatics 21, i369–i377. doi:10.1093/bioinformatics/bti1053

Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., Hume, D.A., 2007. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nat. Rev. Genet. 8, 424–436. doi:10.1038/nrg2026

Sawaya, S., Bagshaw, A., Buschiazzo, E., Kumar, P., Chowdhury, S., Black, M.A., Gemmell, N., 2013. Microsatellite Tandem Repeats Are Abundant in Human Promoters and Are Associated with Regulatory Elements. PLoS ONE 8, e54710. doi:10.1371/journal.pone.0054710

Shemer, J., Adamo, M.L., Roberts, C.T., LeRoith, D., 1992. Tissue-specific transcription start site usage in the leader exons of the rat insulin-like growth

factor-I gene: evidence for differential regulation in the developing kidney. Endocrinology 131, 2793–2799. doi:10.1210/endo.131.6.1446616

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., Weinstock, G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W., Haussler, D., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15, 1034–1050. doi:10.1101/gr.3715005

Sonnenburg, S., Zien, A., Rätsch, G., 2006. ARTS: accurate recognition of transcription starts in human. Bioinformatics 22, e472–e480. doi:10.1093/bioinformatics/btl250

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T.K., Diegel, M., Dunn, D., Ebersol, A.K., Frum, T., Giste, E., Johnson, A.K., Johnson, E.M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E.D., Qu, H., Reynolds, A.P., Roach, V., Safi, A., Sanchez, M.E., Sanyal, A., Shafer, A., Simon, J.M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M.O., Hansen, R.S., Navas, P.A., Stamatoyannopoulos, G., Iyer, V.R., Lieb, J.D., Sunyaev, S.R., Akey, J.M., Sabo, P.J., Kaul, R., Furey, T.S., Dekker, J., Crawford, G.E., Stamatoyannopoulos, J.A., 2012. The accessible chromatin landscape of the human genome. Nature 489, 75–82. doi:10.1038/nature11232

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. Ser. B Methodol. 58, 267–288.

Toffolo, V., Belvedere, P., Colombo, L., Valle, L.D., 2007. Tissue-specific transcriptional initiation of the CYP19 genes in rainbow trout, with analysis of splicing patterns and promoter sequences. Gen. Comp. Endocrinol., Proceedings of the 23rd Conference of European Comparative Endocrinologists: Part 2 153, 311–319. doi:10.1016/j.ygcen.2007.02.013

Villar, D., Flicek, P., Odom, D.T., 2014. Evolution of transcription factor binding in metazoans — mechanisms and functional implications. Nat. Rev. Genet. 15, 221–233. doi:10.1038/nrg3481

Vinh, L.T., Lee, S., Park, Y.-T., D'Auriol, B.J., 2012. A Novel Feature Selection Method Based on Normalized Mutual Information. Appl. Intell. 37, 100–120. doi:10.1007/s10489-011-0315-y

Wakaguri, H., Yamashita, R., Suzuki, Y., Sugano, S., Nakai, K., 2008. DBTSS: database of transcription start sites, progress report 2008. Nucleic Acids Res. 36, D97–D101. doi:10.1093/nar/gkm901

White, N.L., Higgins, C.F., Trezise, A.E., 1998. Tissue-specific in vivo transcription start sites of the human and murine cystic fibrosis genes. Hum. Mol. Genet. 7, 363–369.

Wingender, E., 2008. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. Brief. Bioinform. 9, 326–332. doi:10.1093/bib/bbn016

Xing, E.P., Jordan, M.I., Karp, R.M., 2001. Feature selection for high-dimensional genomic microarray data, in: In Proceedings of the Eighteenth International Conference on Machine Learning. Morgan Kaufmann, pp. 601–608.

Yang, W., Ni, L., Silveyra, P., Wang, G., Noutsios, G.T., Singh, A., DiAngelo, S.L., Sanusi, O., Raval, M., Floros, J., 2013. Motifs within the CA-repeat-rich region of Surfactant Protein B (SFTPB) intron 4 differentially affect mRNA splicing. J. Mol. Biochem. 2, 40–55.

Yuan, Y., Guo, L., Shen, L., Liu, J.S., 2007. Predicting Gene Expression from Sequence: A Reexamination. PLoS Comput Biol 3, e243. doi:10.1371/journal.pcbi.0030243

Yu, L., 2004. Redundancy based feature selection for microarray data, in: In Proc. of SIGKDD. ACM Press, pp. 737–742.

Yu, L., Liu, H., 2004. Efficient Feature Selection via Analysis of Relevance and Redundancy. J Mach Learn Res 5, 1205–1224.

Zhao, X., Xuan, Z., Zhang, M.Q., 2007. Boosting with stumps for predicting transcription start sites. Genome Biol. 8, R17. doi:10.1186/gb-2007-8-2-r17

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67, 301–320. doi:10.1111/j.1467-9868.2005.00503.x

Zou, H., Hastie, T., 2007. Model Building and Feature Selection with Genomic Data, in: Computational Methods of Feature Selection. CRC Press, Boca Raton, FL, pp. 393–411.

Appendices

Appendix A: Review of Feature Reduction Techniques

Information gain (or mutual information) computes the mutual dependence of two variables, or the information that two variables share. This can be used to select the most informative features in a dataset by computed the mutual information between each predictor and the class label. Let H(D) be the entropy of dataset D, and let a be a predictor for the dataset. Information gain is defined as:

$$IG(D, a) = H(D) - H(D|a)$$

This can be computed in parallel for each predictor in the dataset by computing:

$$IG(D, a) = H(D) - \sum_{v \in vals(a)} \frac{|\{x \in D | x_a = v\}|}{|D|} \times H(\{x \in D | x_a = v\})$$

In addition to standard information gain, we also applied the technique proposed in Ding and Peng (2005), a method called MRMR (minimum redundancy, maximum relevance). The algorithm works as follows:

1. Select top-ranked feature by mutual information:

$$i = \underset{i}{\operatorname{argmax}}\, I(c, i)$$

2. Define the following sets:

$$S = \{i\}$$

$$\Omega = \{All\ Features\}$$

$$\Omega_S = \Omega - S$$

3. For each remaining feature:

a. We would like to select the feature which optimizes the following two equations, selecting the maximally-informative feature (with respect to the class label) which is minimally-redundant with already selected features in set *S*:

$$\max_{i \in \Omega_S} I(c, i)$$

$$\min_{i \in \Omega_S} \frac{1}{|S|} \sum_{j \in S} I(i, j)$$

MRMR computes the best feature *i* and adds the feature to the set of selected features *S*:

$$i = \underset{i \in \Omega_S}{\operatorname{argmax}} I(c, i) - \frac{1}{|S|} \sum_{j \in S} I(i, j)$$

$$S = \{S \cup i\}$$

This procedure performs better than information gain in datasets with high redundancy between features by selecting the most informative features (with regards to the class label) which are not explained by other features in the dataset.

Appendix B: Review of Regularization Techniques

In this section, we review several common regularization techniques and the benefits and issues of each. L1-regularization (also known as lasso regularization) attempts to limit the number of features included in a model by imposing a penalty for each included feature (Tibshirani, 1996). This discourages lowly-weighted features which do not make large contributions to the model's performance from being included. L1-regularized logistic regression can be formulated as follows:

$$\min \sum_{i=1}^{M} -\log p\left(y^i \middle| x^i; \theta\right) + \lambda \|\theta\|_1$$

λ is the penalty assigned to the L1 norm of the feature weight vector θ. The L1 norm yields a sparse solution vector, as it aggressively drives small weights towards 0, thereby finding a solution which includes only features which significantly improve performance. In addition, this solution is interpretable, as the importance of a feature to the model is proportional to its feature weight, as opposed to the feature being used to "cancel out" another redundant and correlated feature.

While L1-regularized solution provides *an* optional solution, it is not guaranteed to be unique (the lasso penalty is convex, but not strictly convex). In other words, there can be multiple optimal solutions which are composed of different weightings of different features, essentially forming equivalence classes of features which are equally predictive of the training dataset. For example, if 2 features are perfectly correlated, either feature can be included in the final solution to achieve optimal performance,

however L1-regularization would penalize a model where both features were assigned high weights. Typically L1 will assign a high weight to one feature and a low (or zero) weight to the other. In a classification-only model, this is usually not important, as the goal is to build an accurate, efficient predictor. However, when the interpretation of the model is important, knowing of alternate, equally-performing feature weight assignments is critical. Perhaps a more appropriate weighting of these features (to maximize model interpretability) would be to assign equal weights to each correlated feature.

Zou and Hastie (2005) note that L1 has several other issues when applied to datasets with a large number of predictors. Given *n* training examples and *p* predictors, the lasso can select no more than *n* features in the model. In the reverse case ($p < n$), L1 typically has lower performance than other regularization methods if there is high correlation between many of the predictors (Tibshirani, 1996; Zou and Hastie, 2005). Other regularization methods have been proposed to address these issue, particularly within the context of large-scale genomics data. Zou and Hastie (2007) review several such methods, including elastic net regularized logistic regression, elastic net penalized SVMs, and sparse PCA, a modification of PCA which uses regularization techniques to impose a penalty on non-zero loadings in the principal components.

L2-regularized regression (Hoerl and Kennard, 1988) is another popular regularization method for logistic regression (often called ridge regression). In this method, a bound is placed on the L2-norm of the coefficients, and the residual sum of squares is

minimized within this bound. While the L2-norm bound increases performance over traditional OLS due to its bias-variance tradeoff, ridge regression cannot perform variable selection because no coefficients can be set exactly to 0. This means that correlated variables will not be completely removed from a model like with the lasso, but makes no guarantees about how correlated features will be handled. In addition, uninformative variables will also be included in the model, yielding a model which is less interpretable.

The elastic net regularization and variable selection method introduced by Zou and Hastie (2005) aims to resolve both of these issues by performing variable selection and capturing groups of related variables through the use of the *elastic net penalty*. This penalty can be thought of as a weighted combination of the ridge and lasso penalties. In our results, the mixture of L1 and L2 is controlled by the α parameter, where $\alpha=0$ is a ridge-regularized model and $\alpha=1$ is lasso-regularized model.