# Transcription start site classification

Max Libbrecht, Matt Fisher, Roy Frostig, Hrysoula Papadakis, Anshul Kundaje, Serafim Batzoglou

December 11, 2009

**Abstract**

Understanding the mechanisms of gene expression is a central problem in biology. Important to gene expression is the binding of RNA polymerase at transcription start sites (TSSs). TSS detection from sequence is a well-studied problem, but recent innovations in data collection have made a much richer feature set available. Furthermore, these innovations allow for the large-scale study of TSS activity, which had not previously been affordable. We developed a machine learning model to detect TSSs as well as predict their activity, using newly collected epigenetic data from the ENCODE project. To improve interpretability of our results, we used a learning model that produces a classifer easily understood by a human. Our results improved upon the state of the art in TSS detection, and achieved very high accuracy in predicting TSS activity.

## 1 Introduction

TSS prediction methods based on genomic sequence are widely used [Down and Hubbard, 2002]. The most well-known such method is based on the so-called "TATA box," an AT-rich region which indicates polymerase binding nearby. While these methods can be used to to locate TSSs in a small target area (for example, the region just upstream from a known gene), they generally produce consensus sequences of less than 15 base pairs, and therefore have poor sensitivity/specificity when applied across the entire genome. We propose a machine learning model with a larger feature set in the hope of producing a more discriminative TSS model, using a newly available data set taken by the ENCODE project.

New data collection methods also make it possible to assay for TSS activity on a large scale. We use available such assays to build a machine learning predictor of TSS activity. This approach has not been attempted in the past, because TSS activity data simply was not available.

TSS activity can be broken down into three classes. We call expressed TSSs "active" and non-expressed TSSs "inactive." It is also common for PolII to bind to a TSS, but simply stall instead of transcribing [Muse et al., 2007]. We call this class "poised."

We attempt to distinguish between the four classes non-TSS, inactive, poised and active by building classifiers for three problems: TSS vs. non-TSS, Inactive vs. PolII bound, and Active vs. Poised.

## 2 Learning model

### 2.1 Learning algorithm

Our dataset and our goals have many characteristics that suggest a specific class of learning models. First, initial experiments suggest that there will be many examples that are extremely hard to classify correctly; many TSS's exhibit essentially no features observed in our dataset that characterize it as a TSS. However, there are subsets of TSS's with very identifying features on which we expect very good classification. This motivates a learning algorithm that gives good confidence bounds, enabling us to perform good specificity and sensitivity queries. Second, initial experiments also suggested that the number of training examples available (at least 15000 per

class for all problems) far exceeds the number necessary for good learning. Finally, our input feature vectors are not in the same coordinate system; some represent the response of the genome to various signal tests, while others represent the presence or absence of genomic sequences. While we could attempt to normalize similar elements of the vector to avoid this problem, this could affect the learning algorithm. All three of these characteristics discourage using SVM's—they partition the dataset into only one axis (the margin), which does not necessarily result in good confidence bounds, they are hard to train on a very large dataset, and they are not invariant to coordinate scaling. Decision trees, on the other hand, partition the dataset into similar regions where the class probabilities can be read directly (and thus have good confidence bounds), are easy to train, and are invariant to coordinate scaling. The internals of the classifier they produce are also an intuitive and understandable classifier.

We use a greedy ID3 algorithm to grow the decision tree, with the class probability being read directly from the ratio of the training set in each leaf node. Since the confidence value is very important to us, we seek to avoid false high confidence readings, which are very common in leaves which only contain a small number of examples. Although various techniques can be used to overcome this problem, because the number of training examples available is so large, we use a very simple approach—decisions splits which would produce a node with less than 250 examples are not allowed (cross validation was used to select this number.) We varied numerous parameters, such as the maximum depth allowed, and most reasonable ranges yielded a learner with very comparable performance. For comparison, we also tried a variety of other common machine learning techniques, such as boosted decision stumps, and also noted performance similar to the results shown in this paper. While it was possible with any of these algorithms to drive the training error to zero, this never resulted in good performance; learners with low training error all appeared to be training on the feature noise, which is sensible given our claim that many TSS's do not exhibit any features observed by our dataset.

## 2.2 Feature set

We input the following features into our learning algorithm. We used data taken by the ENCODE project on the K562 cell line. See Section 4: Methods for details on how we generated our training set and how the features were generated from raw data.

1) **Sequence k-mers.** We made one feature for each possible $k$-mer for $k$=4,5,6,7. We extracted all the $k$-mers from a 200bp window centered on each training example.

2) **Histone modifications.** Histone tails are accessible from outside packed chromatin, and can be covalently modified by other proteins. These marks are implicated in gene regulation [Benevolenskaya, 2007, Barski et al., 2007, Koch et al., 2007]. We have data from the following nine histone marks: H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me1, H3K27ac, H3K27me3, H3K36me3, H4K20me1. We aggregated signal from a 1000bp wide window centered on the TSS.

3) **Transcription factors.** Transcription factors are known to play a central role in gene regulation. We have ChIP-seq data for two transcription factors, TAFII and GABP. We aggregated signal from a 1000bp wide window centered on the TSS.

4) **DNA accessibility.** Tightly packed chromatin is inaccessible to transcription machinery. We have data from three different assays measuring DNA accessibility. We aggregated signal from a 1000bp wide window centered on the TSS.

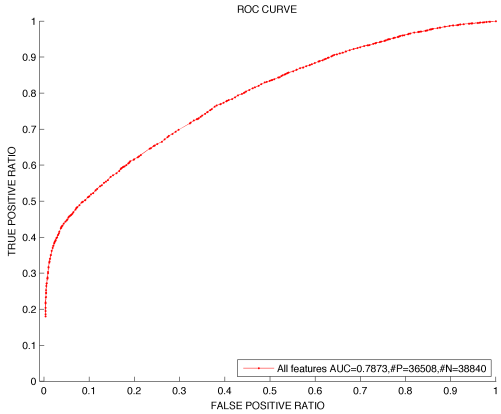
## 2.3 Training set

Our training set consisted of 74627 TSS examples and 79118 non-TSS examples. The TSS examples were further divided into 41578 inactive, 15058 poised and 17991 active examples.

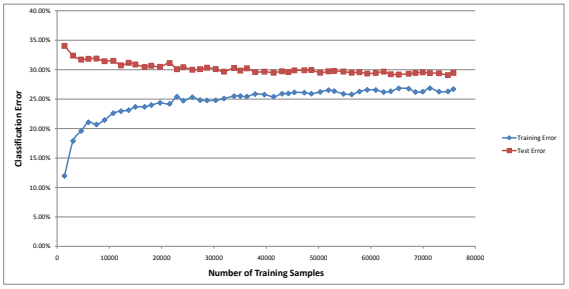
# 3 Results

## 3.1 TSS vs. non-TSS

We achieved 71 percent accuracy on a training hold-out set, which was composed of approximately half positive and half negative examples. We achieved a 78 percent area under an ROC curve (Figure 1). Almost any subset of our features gave comparable accuracy . This suggests that our features are highly correlated.




(a) ROC curve of our classifier's accuracy.




(b) A plot of our test error as a function of number of training examples. The test set in this case was approximately evenly split between positive and negative examples.

Figure 1: Accuracy results on the TSS vs. non-TSS problem.

We improved on the state of the art in TSS detection. Down and Hubbard (2002), who take a sequence-based approach, achieved 54 percent sensitivity and a 74 percent specificity. When matching their sensitivity, we achieve 84 percent specificity and when matching their specificity, we achieve 66 percent sensitivity.

## 3.2 TSS activity

We achieved nearly perfect accuracy in classifying inactive vs. PolII bound (Figure 2). Our classifier picked H3K4me3 as the most important feature, which is supported by the literature.

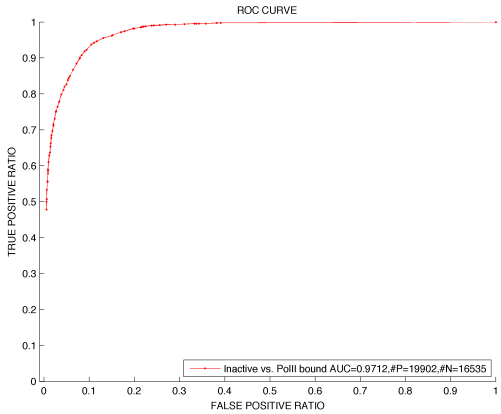
Predicting active vs. poised appears to be a much harder problem. We achieved only 76% area under the ROC curve, compared to significantly higher accuracy on the other two problems. Our algorithm implicates H3K9ac, H3K36me3 and H4K20. All of these features are supported by the literature.

# 4 Methods

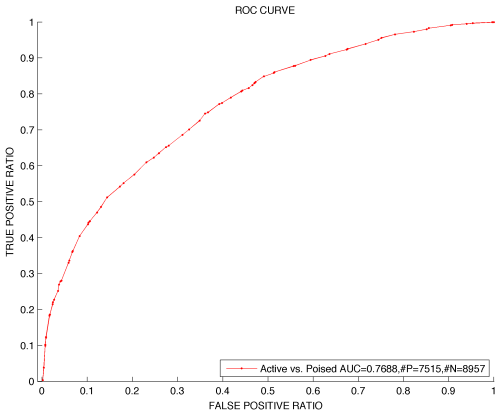
## 4.1 Labels

We used ENCODE TSS annotations as our set of TSS training examples. These were generated by a combination of manual and automatic (unverified) annotation, so some small fraction of these examples may be mis-annotated.

We selected positions at +1000 and –1000 from each TSS as negative examples, throwing out examples less than 999bp from another annotated TSS. Other methods of choosing negative examples, such as selecting random positions in the genome, tend to yield examples that are so dramatically different from true TSSs that they are too easy to classify. By choosing positions near true TSSs, we pick negative examples in a similar genomic




(a) Accuracy in classifying inactive vs. PolII bound




(b) Accuracy in classifying active vs. poised

Figure 2: ROC curves for classifying between active, poised and inactive.

neighborhood, but without the essential character of a true TSS.

We labeled true TSSs as active, poised or inactive as follows. We have ChIP-seq data for PolII binding and RNA-seq data for mRNA expression. We used ENCODE point peak calls on the PolII data and labeled TSSs as bound by PolII if they fell within 400bp of a peak call. We labeled TSSs expressed as mRNA if the 200bp sequence downstream of the TSS has a signal in the RNA-seq experiment greater than 5 RPKM.[1] We labeled those TSSs with both PolII binding and mRNA expression as active, those with PolII binding but no mRNA expression as poised, and those with neither as inactive.

### 4.2 Feature generation

The data for the histone modifications, transcription factors, and DNA accessibility features were taken by the ENCODE project. The data comes from a set of ChIP-seq experiments and the values correspond to unitless signals. The features were generated from the data as an average of the signal over a 1000bp window centered on the TSS. We had an additional feature which corresponded to an increase in the signal from one side of the TSS to the other. This was computed by taking the average of the signal upstream of the TSS, subtracting the average downstream of the TSS, and dividing by the average in the whole window. This feature was not picked by the classifier as informative, which we believe is due to artifacts generated by our implementation. We are working on removing these artifacts, and we expect this feature to become much more informative once this is done.

## 5 Future work

We are currently working on the following improvements to our method.

### 5.1 Additional features

We plan to expand our feature set by adding the following features:

1) **GC content** High GC content is known to correlate with functional elements [Pozzoli et al., 2008].[2]

[1] RPKM is a measure of signal in a sequencing assay, calculated $R10^9/LN$, where $R$ is the number of reads in the window, $L$ is the length of the window, and $N$ is the number of reads in the whole assay.

[2] We've already processed the features for GC content and CpG islands, but the results weren't ready at the time of this writing.

2) **CpG islands** For chemical reasons, CG dinucleotides mutate much faster than average, so CG dinucleotides are relatively rare in the genome except where they are under selective pressure. Regions with high frequencies of CG dinucleotides (CpG islands) are known to correlate with functional elements, in particular, promoters [Saxonov et al., 2006].

3) **Sequence conservation.** Conservation of sequence across different species is suggestive of function.

4) **DNA helix structure** While the 3D structure of the DNA backbone always roughly takes the form of a double-helix, it is modified by the sequence of bases. In particular, the structure of any substrand of 4 bases is determined by the base pair sequence [Packer et al., 2000]. This structure is captured by a measure called the hydroxyl radical cleavage pattern [Price and Tullius, 1993].

5) **Spectrum kernel** By replacing sequence with $k$-mers, we lost a great deal of information. We plan to replace this with running SVMs with a spectrum kernel [Leslie et al., 2002], which captures more of the complexity of the sequence.

### 5.2 TATA box

The so-called TATA box is a well-known indicator of promotor sequences, and several transcription factors are known to bind to its sequence. We expected it would be chosen by the machine learning algorithm as an important $k$-mer feature, but it wasn't, even when we restricted our feature set to only sequence features. This is puzzling, so we plan to investigate it by testing the sequence surrounding each TSS for TATA enrichment by comparing to the TATA consensus sequence. With this, we can analyze the predictive power of the TATA box and hopefully uncover why it isn't selected by our algorithm.

### 5.3 Incorrectly classified TSS/non-TSS examples

As shown in the Results section, on the TSS vs. non-TSS problem, our classification accuracy using small subsets of our feature set is almost as good as it is using the whole set. Despite this, our accuracy is far from perfect. This suggests that while most of our examples are characterized by almost every feature, a fraction of them aren't characterized by any of our features. We plan to analyze this set to see if it is dominated by a particular gene type, chromosome, or other characteristic.

## References


[Barski et al., 2007] Barski, A., Cuddapah, S., Cui, K., Roh, T., Schones, D., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837.

[Benevolenskaya, 2007] Benevolenskaya, E. (2007). Histone H3K4 demethylases are essential in development and differentiation. *Biochemistry and Cell Biology*, 85(4):435–443.

[Down and Hubbard, 2002] Down, T. and Hubbard, T. (2002). Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Research*, 12(3):458.

[Koch et al., 2007] Koch, C., Andrews, R., Flicek, P., Dillon, S., Kara "oz, U., Clelland, G., Wilcox, S., Beare, D., Fowler, J., Couttet, P., et al. (2007). The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome research*, 17(6):691.

[Leslie et al., 2002] Leslie, C., Eskin, E., and Noble, W. (2002). The spectrum kernel: A string kernel for SVM protein classification. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 7, pages 566–575.

[Muse et al., 2007] Muse, G., Gilchrist, D., Nechaev, S., Shah, R., Parker, J., Grissom, S., Zeitlinger, J., and Adelman, K. (2007). RNA polymerase is poised for activation across the genome. *Nature genetics*, 39(12):1507.

[Packer et al., 2000] Packer, M., Dauncey, M., and Hunter, C. (2000). Sequence-dependent DNA structure: tetranucleotide conformational maps. *Journal of molecular biology*, 295(1):85–103.

[Pozzoli et al., 2008] Pozzoli, U., Menozzi, G., Fumagalli, M., Cereda, M., Comi, G., Cagliani, R., Bresolin, N., and Sironi, M. (2008). Both selective and neutral processes drive gc content evolution in the human genome. *BMC Evolutionary Biology*, 8(1):99.

[Price and Tullius, 1993] Price, M. and Tullius, T. (1993). How the structure of an adenine tract depends on sequence context: a new model for the structure of TnAn DNA sequences. *Biochemistry*, 32(1):127–136.

[Saxonov et al., 2006] Saxonov, S., Berg, P., and Brutlag, D. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences*, 103(5):1412.