Classification of CAGE Peaks

- TSS vs non-TSS

by Timo Lassmann (timolassmann at gmail dot com)

March 27, 2014

Method

A training set comprised of both positive and negative sequences is extracted from the data. Gaussian mixture models are trained to capture the relative distribution of 4-mer occurrences surrounding TSSs. Each sequence is scored against all models resulting in a 256 vector of values for each sequence. The latter together with the cluster label is used to derive a random decision tree ensemble model. Finally, the RDT model is used to classify test sequences not used in the training of any models. The entire procedure is repeated many times and predictions for each cluster averaged (see Fig. 1).

Data and Parameters

We used the permissive DPI clusters as the basis for our predictions. All clusters within 100bp from a known transcriptional start site were labelled as positive, remaining sequences as negative. We extracted sequences centered on the middle of each CAGE cluster of various lengths (100bp ,200bp, 400bp, 600bp, 800bp, 1kb, 2kb). Increasing the number of mixtures had little effect on the prediction accuracy. We therefore ran our predictor using only a single gaussian. In each run we ran 2,4,6,8,10-fold cross validation, each one 5 times. For each cluster the prediction scores were averaged over all runs in which the cluster was not used for training the model.

In addition we ran the same procedure using ncRNAs from gencode v13 as a standard of truth.

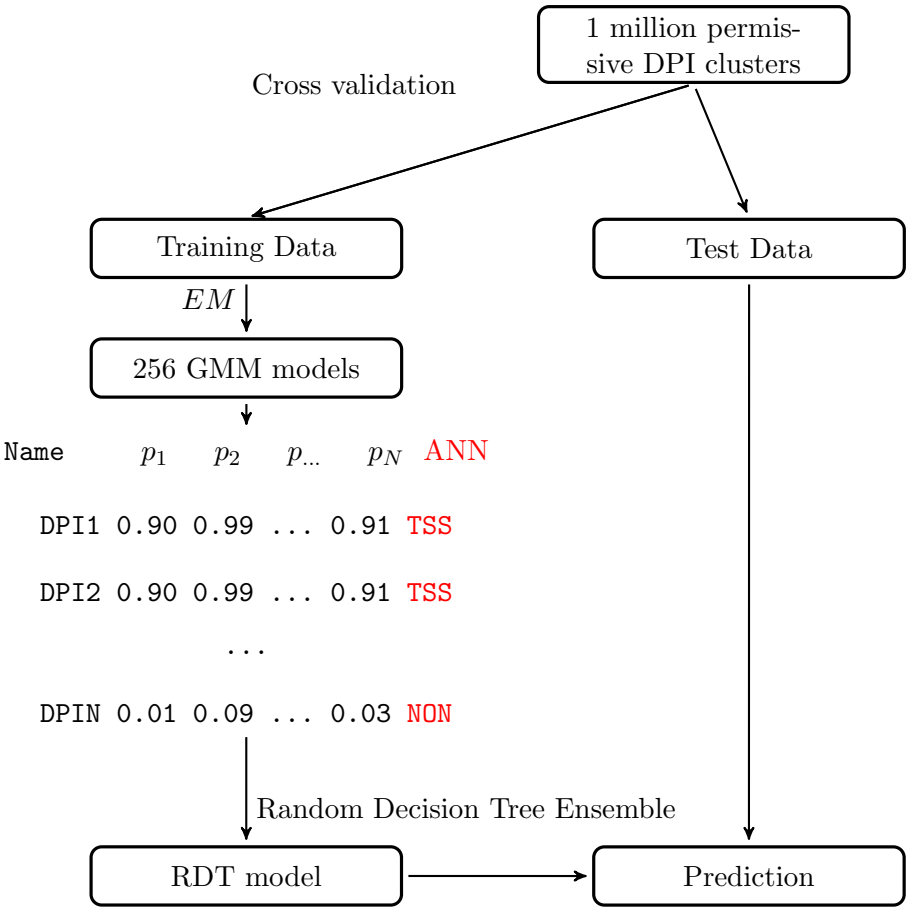

Figure 1: **Algorithm workflow.**

Results

We used ROC curves to assess the accuracy of our classifier. All novel TSS clusters are counted as false positives making this assessment very strict. When increasing the window surrounding clusters the prediction accuracy increases (Fig. 2). We also attempted to run our predictor using solely upstream sequences. As smaller window sizes this predictor performs a little worse. It is worthwhile noting that just using a 100bp window a reasonable AUC of 0.75 can be achieved. For the remainder of our analysis we used the 2kb window setting.

Compared to known gene models our methods achieved an AUC of 0.93; compared to TSS models derived from segmentation of ENCODE histone modification tracks 0.83 (Figure 3,5). We used these curves to derive 2 thresholds on our prediction scores (shown in the plots). In mouse, ENCODE predictions are currently unavailable and we only compared against known gene models (Fig. 6).

We evaluated the performance of our main predictor on only ncRNAs (Fig. 4). The accuracy is lower than on all known TSSs but still reasonable. In addition we constructed a dedicated ncRNA TSS classifier (Fig. 7).

Files

For human we provide one bed files: `TSS_human.bed`. Clusters are separated into three classes: non-TSS (grey), TSS based on a relaxed 0.14 threshold (blue) and strict TSS predictions based on a 0.228 threshold (green). In mouse we only provide TSS and non-TSS classes: `TSS_mouse.bed`.

The fourth field in the file lists the name of the peak followed by a comma and the score assigned by the TSS classifier. The fifth field is the distance of the peak to the nearest annotated TSS.

Source Code

The source code to perform this analysis is part go the tomeTools package(http://tometools.sourceforge.net).

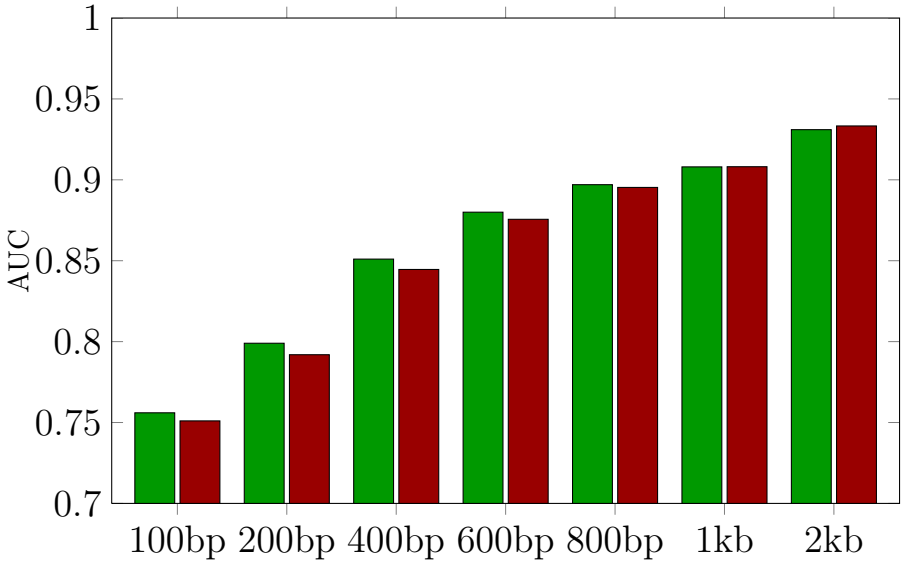

Figure 2: **Effect of sequence length on prediction accuracy.** Prediction were made using sequences surrounding DPI clusters (green) or just upstream sequences (red).

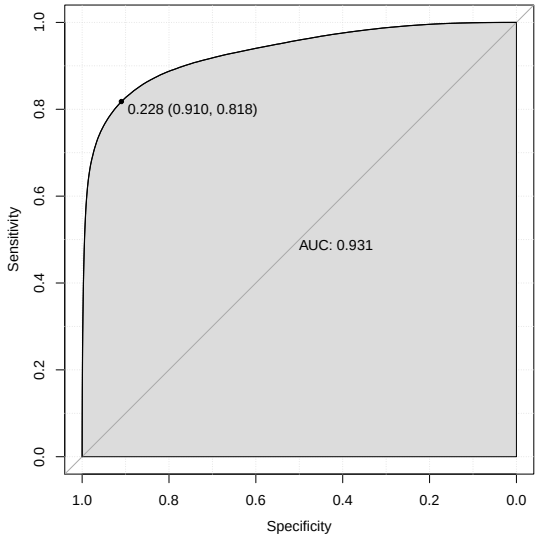

Figure 3: **ROC curve demonstrating the agreement of TSS prediction with known promoter regions (in human).** As the standard of truth we used DPI clusters within 100bp of known models.

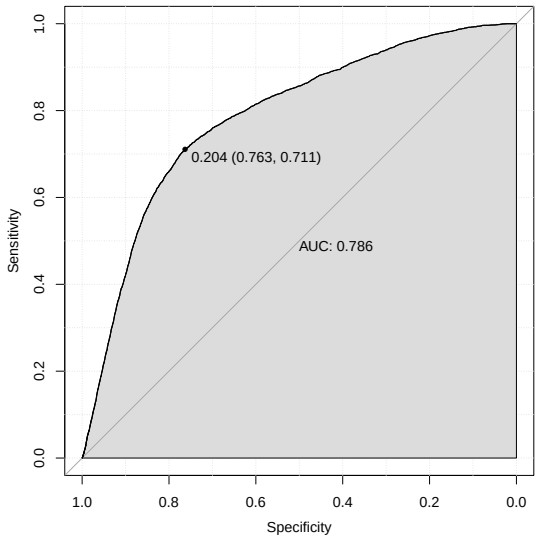

Figure 4: **ROC curve demonstrating the agreement of TSS prediction with known ncRNA promoter regions (in human).** As the standard of truth we used DPI clusters within 100bp of known gencode v13 non-coding and linc RNA transcripts.

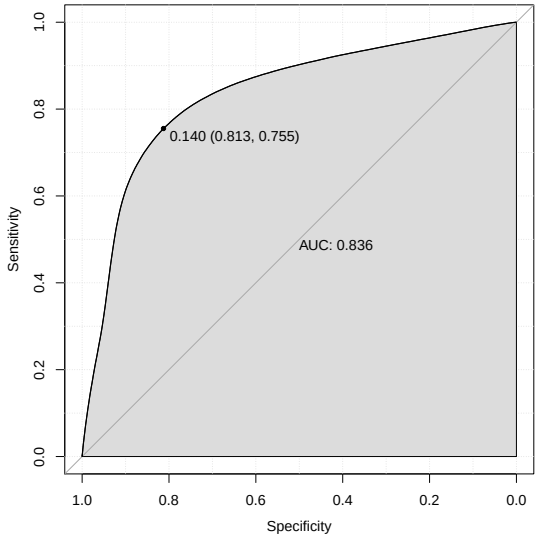

Figure 5: **ROC curve demonstrating the agreement of TSS prediction with ENCODE TSS prediction.** As the standard of truth we used all regions labelled as active, weak and poised promoter in any of the ENCODE cell lines (GM12878, H1hesc, Hepg2, Hmec, Hsmm, Huvec, K562, Nhekm, Nhlf).

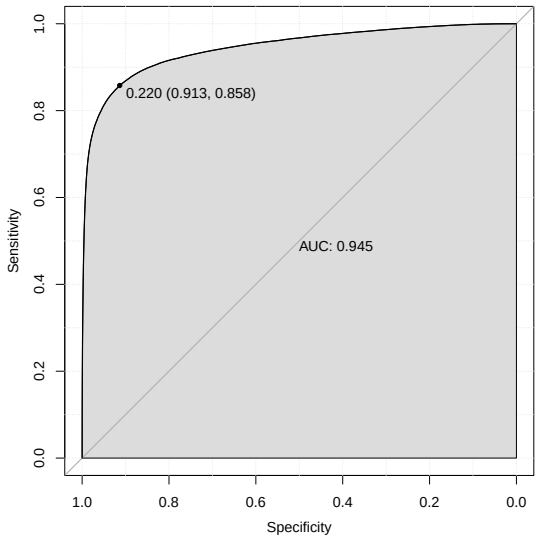

Figure 6: **ROC curve demonstrating the agreement of TSS prediction with known promoter regions (in mouse).** As the standard of truth we used DPI clusters within 100bp of known models.

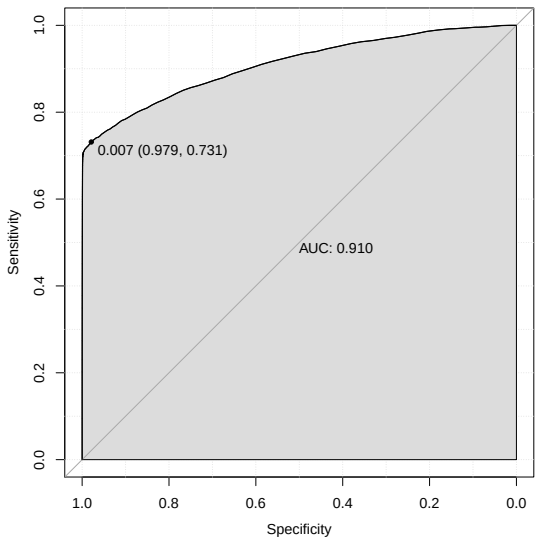

Figure 7: **ROC curve demonstrating the agreement of a ncRNA TSS predictor with known ncRNA promoter regions (in human).** As the standard of truth we used DPI clusters within 100bp of known gencode v13 non-coding and linc RNA transcripts.

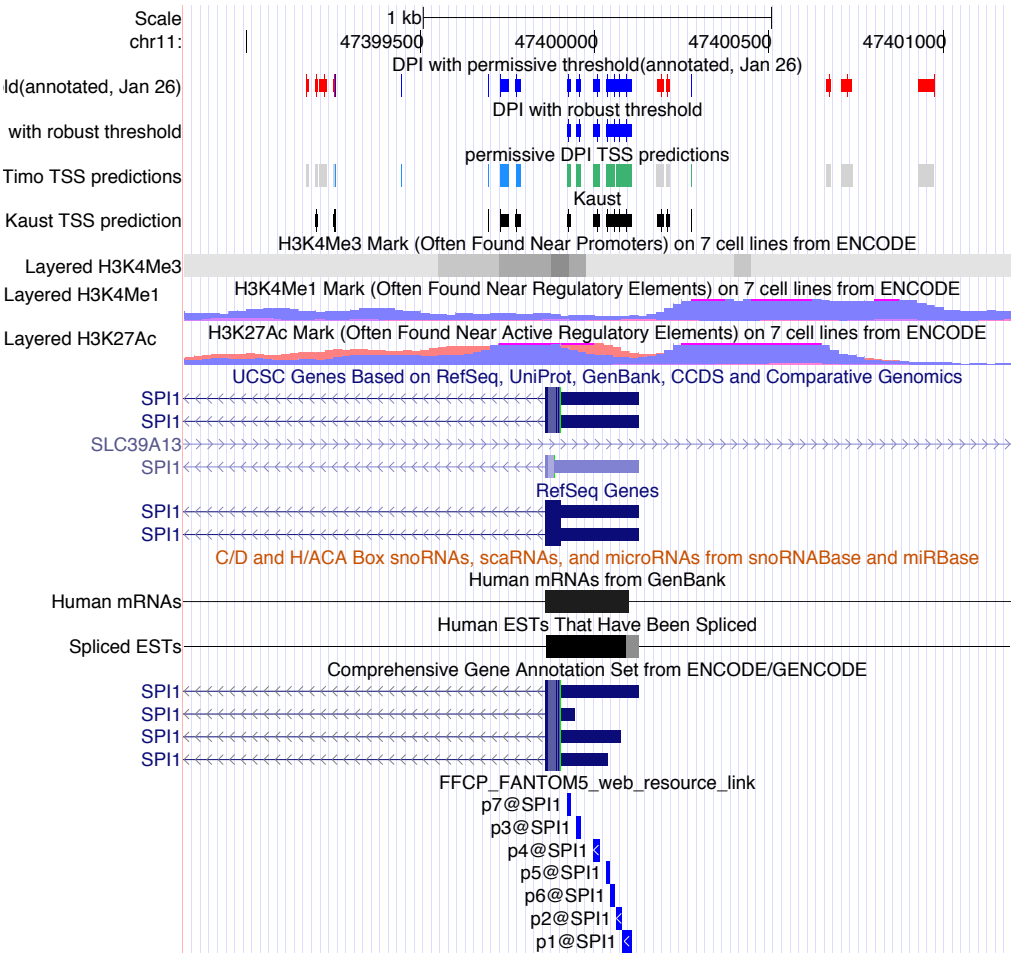

Figure 8: **Screenshop of the SPi1 loci.**

Bibliography

[1] Kanamori-Katayama, M. et. al. Unamplified cap analysis of gene expression on a single-molecule sequencer. Genome Res (2011) vol. 21 pp. 1150-1159