k-mer analysis of EnhancerSeq

Anurag Sethi
P2

EnhancerSeq (preliminary)

Genome

DNase
1/7 of DNaseSeq peaks

9000 regions (approx. 500 bp)

EnhancerSeq

75% of regions tested

2000 peaks
(30% have enhancer activity)

1396 regions
Overlapping with our predictions

722 peaks
(52% have enhancer activity)

MCF-7 cell line - Matched filter on H3K27ac predictions - 30,407 predictions (5% FDR) - vary in length (350-1500 bp typically).

Eventual goal of our predictions for ENCODE Cancer

30407 predictions (average of 2.1 kb) - Matched Filter

Sequence-based Bayesian strategy

10000 predictions (500 bp)

Can we prioritize 10000 predictions that have the best chance of being positive in the EnhancerSeq assay?

Use the sequences of the positives in current EnhancerSeq assay to improve our predictions

Create a naive Bayes model for a particular sequence being an enhancer

$$P(+|\text{sequence},MF) = \frac{P(+|MF)\ P(\text{sequence}|+,MF)}{P(\text{sequence}|MF)}$$

The assumption is the sequence can be decomposed into independent k-mers (big assumption)

$$P(+|\text{counts},MF) = \frac{P(+|MF)\ P(\text{counts}|+,MF)}{P(\text{sequence}|MF)}$$

Start with Naive Bayes Model

$$P(+|\text{counts},MF) \approx P(+|MF)\, P(c_1, c_2, \ldots.|+,MF)$$

$$\approx P(+|MF)\, \Pi_i P(c_i| +,MF)$$

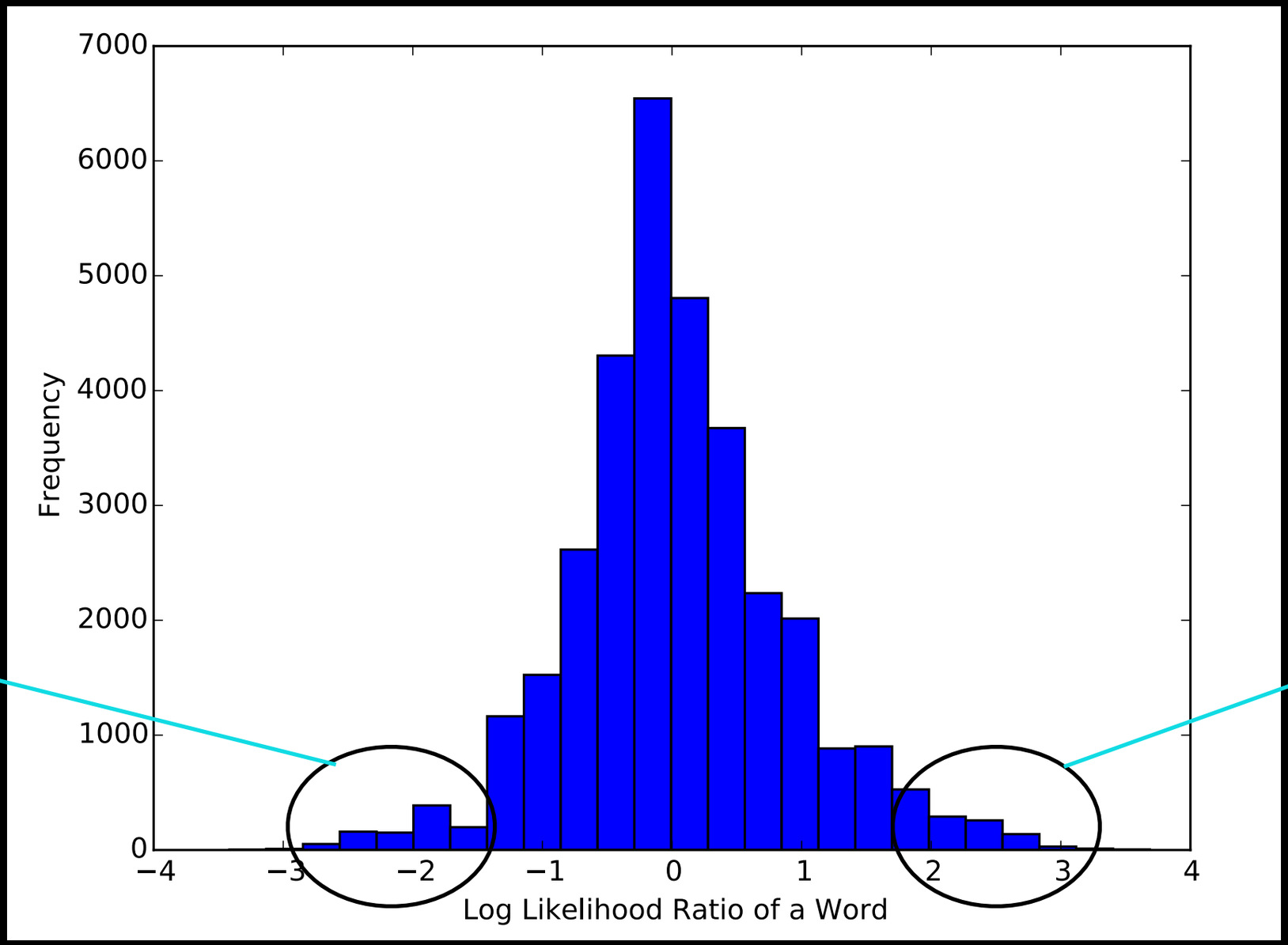
Likelihood of a positive from MF being a positive versus a negative

$$\frac{P(+|\text{counts},MF)}{P(-|\text{counts},MF)} = \frac{P(+|MF)\, P(\text{counts}|+,MF)}{P(-|MF)\, P(\text{counts}|-,MF)}$$

$$\frac{P(+|\text{counts},MF)}{P(-|\text{counts},MF)} = \frac{P(+|MF)\, P(\text{counts}|+,MF)}{P(-|MF)\, P(\text{counts}|-,MF)}$$

Enrichment of 8-mer words in positives over negatives

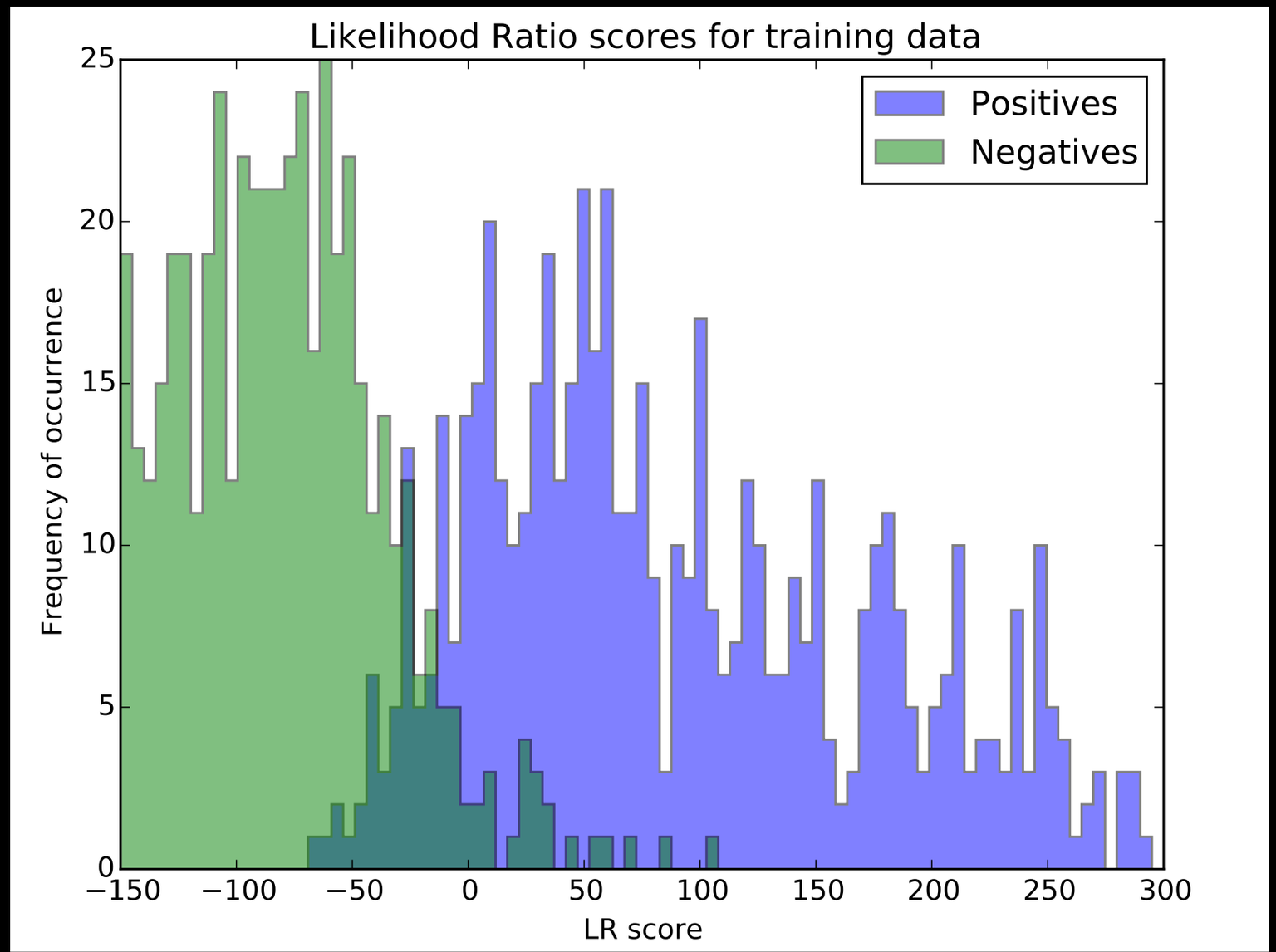
Depleted in positive



Enriched in positive - Motif hunting

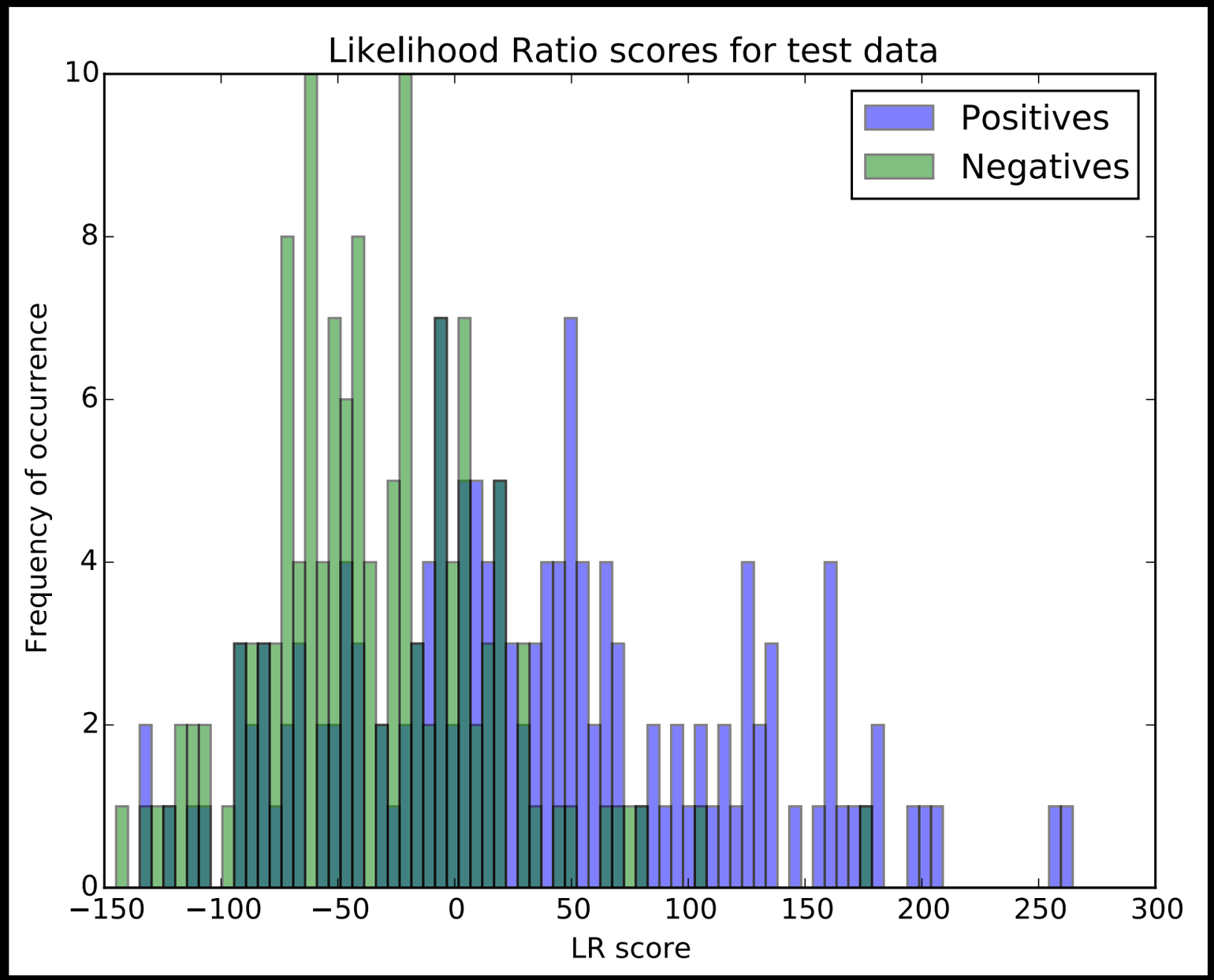
Identification of 8-mers enriched and depleted in positives over negatives

Likelihood of training sequence based upon 8-mers



Training data gets well separated

Likelihood of test sequence based upon 8-mers



However the story is not that simple

Performance of model on test data