

1 Introduction

The emergence of ChIP-seq technology for genome-wide profiling of transcription factor binding sites (TFBS) has made it possible to categorize very precisely the TFBS motifs. How to harness the power of huge volume of data generated by this new technology presents many computational challenges. We propose a novel motif discovery algorithm that is scalable to large databases, and performs discriminative motif discovery by searching the most differential motifs between a foreground and background sequence dataset. This tool can be used in a traditional setting in which the foreground sequence dataset is derived from a ChIP-seq binding profile, and background sequence dataset is either sampled from the genome or generated from a null model. It can also be used for comparative study involving two TFBS binding profiles.

In a nutshell, the method works as the following: we enumerate all fixed-length n-mers exhaustively, and measure their discriminative power by a logistic regression model. The top ranking seed motif is then iteratively refined by allowing IUPAC degenerate letters and extended to a longer motif automatically. We introduce a bootstrapping robustness test to avoid over-fitting in the optimization process. The logistic regression framework offers direct measurement of statistical significance, and we demonstrate by permutation tests that the z-value statistics do reflect the probability of occurrence by chance. Compared to traditional motif finding tool, use of proper control sequences for comparison avoids the difficulty of modeling true genomic background, which usually presents complicated high order structure such as dinucleotide sequence preference, repeats, nucleosome positions signals, etc. When used to compare two similar ChIP-Seq samples, the discriminative motifs usually leads to insights on sample specificity.

2 Example

We have applied this technique to the CTCF chip-seq experiment. The positive dataset contains 10,000 CTCF chip-seq binding sites, each with 200 bases. The negative dataset contains the same number, and the same length of sequences as the positive set. They are chosen from chip-seq mapped regions with low coverage, and they share the same distribution of distance to transcription start site as the positives to adjust for any promoter bias.

```
> library(motifRG)
> data(ctcf.seq)
> data(control.seq)
> ### concatenate the foreground, background sequences
> all.seq <- append(ctcf.seq, control.seq)
> ### specify which sequences are foreground, background.
> category <- c(rep(1, length(ctcf.seq)), rep(0, length(control.seq)))
```

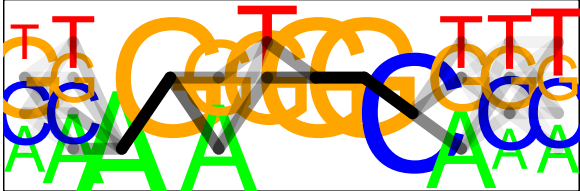
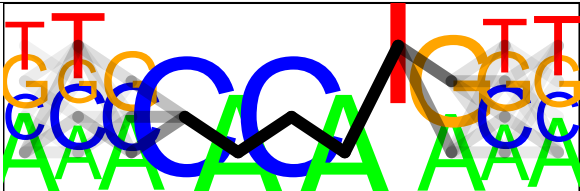
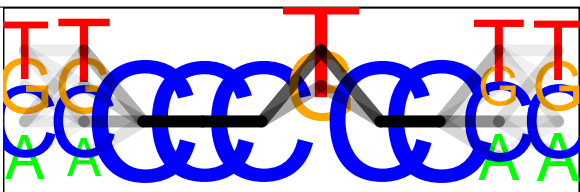
```

> ### find motifs
> ctfc.motifs <- findMotif(all.seq=all.seq, category=category, max.motif=3)

> motifLatexTable(main="CTCF motifs", ctfc.motifs)

```

Table 1: CTCF motifs

Consensus	scores	ratio	fg.frac	bg.frac	logo
NNAGRKGGCDNN	17.4	4.47	0.54	0.13	
NNVCACATRNN	8.6	2.73	0.28	0.14	
NNCCCTCCNN	-8.5	0.39	0.18	0.34	

```

> ### Find a refined PWM model given the motif matches as seed
> pwm.match <- refinePWMMotif(ctfc.motifs$motifs[[1]]@match$pattern, ctfc.seq)
> library(seqLogo)

```

```

> ## Motifs found by findMotif tend to be relatively short, as longer and
> ## more specific motif models do not necessarily provide better
> ## discrimination of foreground background vs background if they are
> ## already well separated. However, one can refine and extend a PWM model
> ## given the motif matches by findMotif as seed for more specific model.
> pwm.match.extend <-
+   refinePWMMotifExtend(ctfc.motifs$motifs[[1]]@match$pattern, ctfc.seq)

```

```
> seqLogo(pwm.match$model$prob)
```

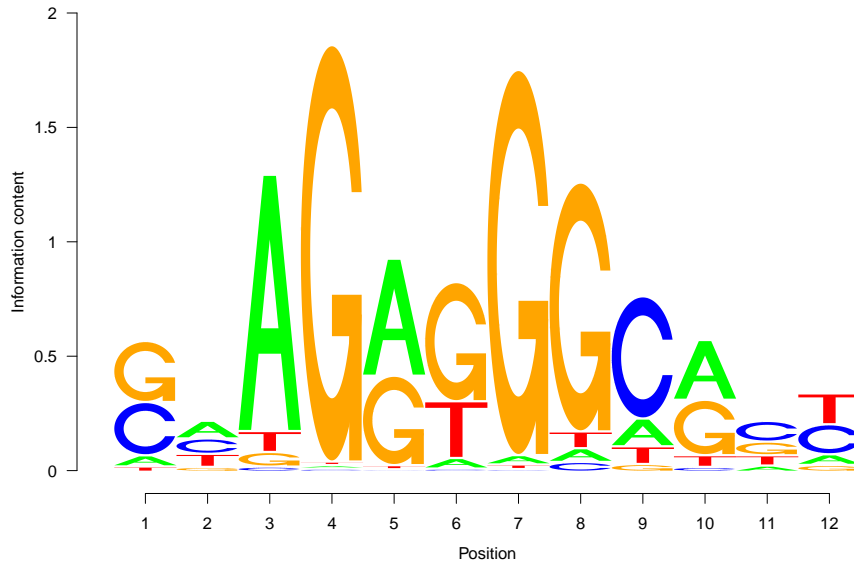


Figure 1: PWM logo of CTCF PWM matches

```
> seqLogo(pwm.match.extend$model$prob)
```

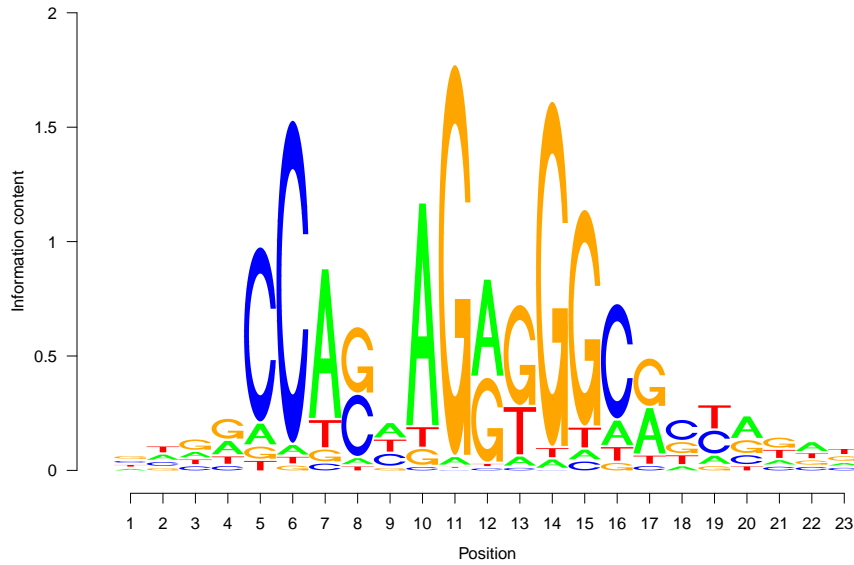


Figure 2: PWM logo of CTCF PWM matches

```
> plotMotif(pwm.match.extend$match$pattern)
```

