

Motifs and Position Weight Matrices

Martin Morgan¹

June 23 – 28, 2013

¹mtmorgan@fhcrc.org

Position weight matrix (PWM)

- ▶ A probabilistic description of (short) sequences
- ▶ Useful for describing, e.g., transcription factor binding sites
- ▶ Easy to construct and use, e.g., scanning genomic sequences for occurrence of binding motifs
- ▶ A quick case study

Position weight matrix

- ▶ HNF4alpha transcription factor binding sites

A DNAMat instance of length 71
width seq

[1] 13 AGTTCAAGGATCA

[2] 13 GGGGTCAAGGGTT

... ..

[71] 13 AAACCAAAGTTCA

- ▶ Consensus matrix / position frequency matrix (first 7 columns)

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]
A	29	2	13	5	3	63	56
C	7	2	5	23	53	1	2
G	30	60	35	20	4	3	11
T	5	7	18	23	11	4	2

Position weight matrix

- ▶ Approximately: log probability p_{ij} of symbol i at position j , given background (prior) probability b_i , $\log_2(p_{ij}/b_i)$
- ▶ Scaled so possible score ranges from 0 to 1

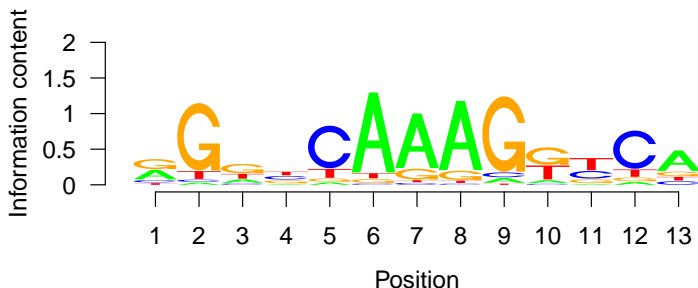
	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]
A	0.065	-0.007	0.042	0.016	0.003	0.087	0.083
C	0.025	-0.007	0.016	0.058	0.082	-0.024	-0.007
G	0.066	0.085	0.070	0.054	0.010	0.003	0.038
T	0.016	0.025	0.051	0.058	0.038	0.010	-0.007

- ▶ Score a 'subject' sequence by adding elements of corresponding entries in PWM

Use

```
library(Biostrings)
data(HNF4alpha)
pfm <- consensusMatrix(HNF4alpha)
pwm <- PWM(consensusMatrix(HNF4alpha))
```

seqLogo:



Use

```
library(BSgenome.Dmelanogaster.UCSC.dm3)
hits <- matchPWM(pwm, Dmelanogaster$chr3R)
length(hits)
```

```
[1] 27164
```

```
head(hits)
```

Views on a 27905053-letter DNAString subject

subject: GAATTCTCTCTTGTGTAG...GTTCGCATTCTAGGAATTC
views:

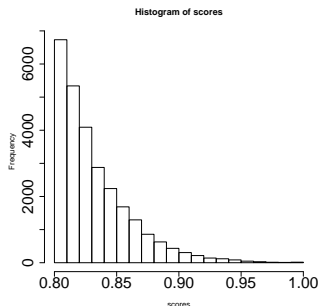
	start	end	width	
[1]	485	497	13	[GGGGTAAAGAGCT]
[2]	1415	1427	13	[GAACCAAAGTTCC]
[3]	6264	6276	13	[GGGTCATAGTTCC]
[4]	7573	7585	13	[TGGCCAACGTTCA]
[5]	7655	7667	13	[GTATCAAAGTGCC]
[6]	8086	8098	13	[CGGCCAAAGCCCG]

Use

► Scores

```
scores <-  
  PWMscoreStartingAt(  
    pwm, subject(hits),  
    start(hits))  
hist(scores)
```

- Minus strand – take the reverseComplement of the PWM!



Software

- ▶ *Biostrings* for PWM matching
- ▶ *MotifDb* for a catalog of known motifs
- ▶ *seqLogo* for visualization

Case study

- ▶ Huang et al., 2013, Highly Recurrent TERT Promoter Mutations in Human Melanoma, Science 339:957
- ▶ Prior whole-genome sequencing of malignant melanomas; two somatic telomerase reverse transcriptase (TERT) promoter mutations in 17 of 19 cases: C228T and C250T
- ▶ Validate & extend to additional tumor cell lines
- ▶ C228T and C250T introduce E-twenty-six (ETS) transcription factor binding sites
- ▶ Suggests plausible MAP kinase pathway; recurrent somatic mutations in regulatory regions as driver events in cancer

Case study

Lab workflow:

- ▶ (Call variants in TERT promoter region)
- ▶ Score catalog of known PWMs for match to reference versus variant sequence

Acknowledgments

- ▶ Paul Shannon (*MotifDb*; lab workflow)
- ▶ Hervé Pagès (*Biostrings*)
- ▶ Patrick Aboyoun (PWM & friends)
- ▶ Valerie Obenchain (lab workflow)