

# Algoritmos para Análise de Sequências Biológicas

Motifs probabilísticos

# Sumário

- Motifs probabilísticos
- PWM e PSSM
- Probabilidade de geração de uma sequência
- Sequência mais provável

# Perfis probabilísticos

- Um **perfil** pode ser representado por uma *Position Weighted Matrix* (PWM)
- Colunas representam as posições do motif
- Linhas representam os possíveis caracteres do alfabeto
- Posições da matriz indicam probabilidade de aparecer o carácter nessa posição
- Estas probabilidades podem ser convertidas em scores usando a mesma estratégia da geração de matrizes de substituição (log odds) – neste caso ficamos com uma *Position Specific Scoring Matrix* (PSSM)

# Geração da PWM

- Efetuar as contagens
- Calcular as frequências
- Podem adicionar-se pseudocontagens para evitar probabilidades de zero na PWM
- PSSM é obtida pelo logaritmo da probabilidade a dividir pela probabilidade esperada  $\frac{1}{n}$  em que  $n$  é o nº de símbolos

# Perfis probabilísticos

HEM13 CCCATTGTTCTC

HEM13 TTTCTGGTTCTC

HEM13 TCAATTGTTTAG

ANB1 CTCATTGTTGTC

ANB1 TCCATTGTTCTC

ANB1 CCTATTGTTCTC

ANB1 TCCATTGTTTCGT

ROX1 CCAATTGTTTTG

Cons YCHATTGTTCTC

A 002700000010

C 464100000505

G 000001800112

T 422087088261

# PWMs

## Sem pseudocontagens

	1	2	3	4	5	6	7	8	9	10	11	12
A	$\frac{0}{8}$	$\frac{0}{8}$	$\frac{2}{8}$	$\frac{7}{8}$	$\frac{0}{8}$	$\frac{0}{8}$	$\frac{0}{8}$	$\frac{0}{8}$	$\frac{0}{8}$	$\frac{0}{8}$	$\frac{1}{8}$	$\frac{0}{8}$
C	$\frac{4}{8}$	$\frac{6}{8}$	$\frac{4}{8}$	$\frac{1}{8}$	$\frac{0}{8}$	$\frac{0}{8}$	$\frac{0}{8}$	$\frac{0}{8}$	$\frac{0}{8}$	$\frac{5}{8}$	$\frac{0}{8}$	$\frac{5}{8}$
G	$\frac{0}{8}$	$\frac{0}{8}$	$\frac{0}{8}$	$\frac{0}{8}$	$\frac{0}{8}$	$\frac{1}{8}$	$\frac{8}{8}$	$\frac{0}{8}$	$\frac{0}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{2}{8}$
T	$\frac{4}{8}$	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{0}{8}$	$\frac{8}{8}$	$\frac{7}{8}$	$\frac{0}{8}$	$\frac{8}{8}$	$\frac{8}{8}$	$\frac{2}{8}$	$\frac{6}{8}$	$\frac{1}{8}$

# PWMs

Com pseudocontagem de 1

	1	2	3	4	5	6	7	8	9	10	11	12
A	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{3}{12}$	$\frac{8}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{1}{12}$
C	$\frac{5}{12}$	$\frac{7}{12}$	$\frac{5}{12}$	$\frac{2}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{6}{12}$	$\frac{1}{12}$	$\frac{6}{12}$
G	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{9}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{2}{12}$	$\frac{3}{12}$
T	$\frac{5}{12}$	$\frac{3}{12}$	$\frac{3}{12}$	$\frac{1}{12}$	$\frac{9}{12}$	$\frac{8}{12}$	$\frac{1}{12}$	$\frac{9}{12}$	$\frac{9}{12}$	$\frac{3}{12}$	$\frac{7}{12}$	$\frac{2}{12}$

# PSSM

$$score = \log_2 \frac{\frac{f+p}{n+b \times p}}{\frac{1}{b}}$$

**f** frequência absoluta

**p** pseudocontagem

**n** nº de sequências

**b** nº de símbolos



# PSSM (com pseudocontagem de 0.5)

A 002700000010  
C 464100000505  
G 000001800112  
T 422087088261

f frequência absoluta

p 0.5

n 8

b 4

A	-2.32	-2.32	0.00	1.58	-2.32	-2.32	-2.32	-2.32	-2.32	-2.32	-0.74	-2.32
C	0.85	1.38	0.85	-0.74	-2.32	-2.32	-2.32	-2.32	-2.32	1.14	-2.32	1.14
G	-2.32	-2.32	-2.32	-2.32	-2.32	-0.74	1.77	-2.32	-2.32	-0.74	-0.74	0.00
T	0.85	0.00	0.00	-2.32	1.77	1.58	-2.32	1.77	1.77	0.00	1.38	-0.74

# Probabilidade de gerar uma sequência

$$P(s|P) = \prod_{i=1}^n p_{s_i,i}$$

- $P(s|P)$  probabilidade da sequência  $s$  ser criada pelo perfil PWM  $P$
- $p_{s_i,i}$  probabilidade de encontrar  $s_i$  na posição  $i$
- Se  $s$  é próxima do consenso então a  $P(s|P)$  é alta
- Se  $s$  é muito diferente do consenso então a  $P(s|P)$  é baixa
- Se usarmos a PSSM, então somamos os scores tal como acontece nos alinhamentos

# Probabilidade de gerar uma sequência

```
>>> P=pwm(['ATTG','ATCG','ATTC','ACTC'], pseudocount = 0.5)
A  0.75  0.08  0.08  0.08
C  0.08  0.25  0.25  0.42
G  0.08  0.08  0.08  0.42
T  0.08  0.58  0.58  0.08
>>> prob_seq("ATTG", pwm = P)
0.10596599999999998
>>> prob_seq("CCCA", pwm = P)
0.0004
>>> 0.75*0.58*0.58*0.42
0.10596599999999998
>>> 0.08*0.25*0.25*0.08
0.0004
```

# Sequência mais provável

Dada uma sequência  $S$ , encontrar a subsequência  $s$  com maior probabilidade de ter sido gerada pelo PWM.

```
>>> for s in re.findall('(?=(...))', "ACCGTGA"):
    print(s, prob_seq(s, pwm = P))
```

```
ACCG 0.01953125
```

```
CCGT 0.00014467592592592592
```

```
CGTG 0.0016878858024691357
```

```
GTGA 0.00033757716049382714
```

```
>>> seq_mais_provavel("ACCGTGA", pwm = P)
['ACCG']
```