Regulatory sequence analysis

Matrix-based pattern matching

Jacques van Helden

https://orcid.org/0000-0002-8799-8584

Aix-Marseille Université, France

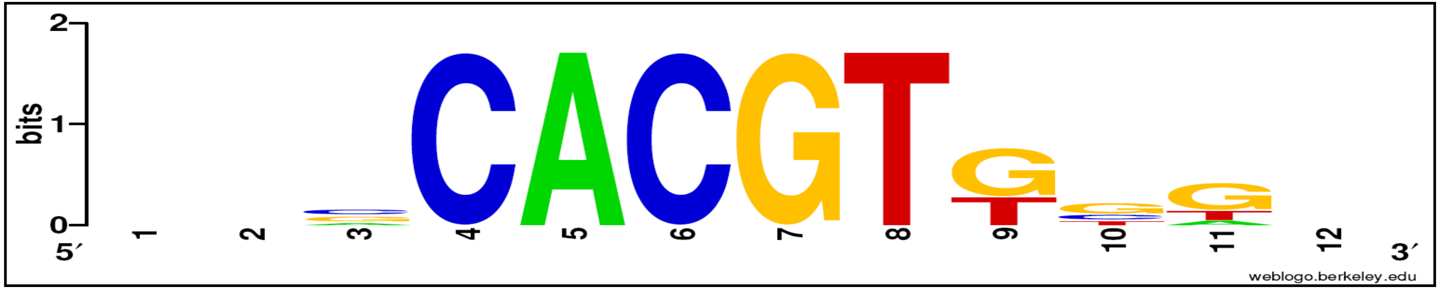
Theory and Approaches of Genome Complexity (TAGC)

Institut Français de Bioinformatique (IFB)

http://www.france-bioinformatique.fr

Regulatory motif : position-specific scoring matrix (PSSM)
Binding motif of the yeast TF Pho4p (TRANSFAC matrix F$PHO4_01)

Pos Base	1	2	3	4	5	6	7	8	9	10	11	12
A	1	3	2	0	8	0	0	0	0	0	1	2
C	2	2	3	8	0	8	0	0	0	2	0	2
G	1	2	3	0	0	0	8	0	5	4	5	2
T	4	1	0	0	0	0	0	8	3	2	2	2
			V	C	A	C	G	T	K	B		



Frequency matrix

Frequency matrix

Residue\position	1	2	3	4	5	6	7	8	9	10	11	12	Prior (p_i)
A	0,125	0,375	0,250	0,000	**1,000**	0,000	0,000	0,000	0,000	0,000	0,125	0,250	0,25
C	0,250	0,250	**0,375**	**1,000**	0,000	**1,000**	0,000	0,000	0,000	0,250	0,000	0,250	0,25
G	0,125	0,250	**0,375**	0,000	0,000	0,000	**1,000**	0,000	**0,625**	**0,500**	**0,625**	0,250	0,25
T	0,500	0,125	0,000	0,000	0,000	0,000	0,000	**1,000**	**0,375**	0,250	0,250	0,250	0,25
Sum	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

$$\boxed{f_{i,j} = \frac{n_{i,j}}{\sum_{i=1}^{A} n_{i,j}}}$$

A	*alphabet size (=4)*
$n_{i,j,}$	*occurrences of residue i at position j*
p_i	*prior residue probability for residue i*
$f_{i,j}$	*relative frequency of residue i at position j*

Count matrix with pseudo-count

1st option: identically distributed pseudo-weight (equiprobable residue priors)

Count matrix with pseudo-count **k= 1** **Equiprobable residues**

Residue\position	1	2	3	4	5	6	7	8	9	10	11	12	Prior (pi)
A	1,25	3,25	2,25	0,25	**8,25**	0,25	0,25	0,25	0,25	0,25	1,25	2,25	0,25
C	2,25	2,25	**3,25**	**8,25**	0,25	**8,25**	0,25	0,25	0,25	2,25	0,25	2,25	0,25
G	1,25	2,25	**3,25**	0,25	0,25	0,25	**8,25**	0,25	**5,25**	**4,25**	**5,25**	2,25	0,25
T	4,25	1,25	0,25	0,25	0,25	0,25	0,25	**8,25**	**3,25**	2,25	2,25	2,25	0,25
Sum	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	1,00

$$f'_{i,j} = \frac{n_{i,j} + k/A}{\sum_{i=1}^{A} n_{i,j} + k}$$

2nd option: pseudo-weights distributed according to residue-specific priors

Count matrix with pseudo-count **k= 1** **Specific nucleotide frequencies**

Residue\position	1	2	3	4	5	6	7	8	9	10	11	12	Prior (pi)
A	1,33	3,33	2,33	0,33	**8,33**	0,33	0,33	0,33	0,33	0,33	1,33	2,33	0,33
C	2,17	2,17	**3,17**	**8,17**	0,17	**8,17**	0,17	0,17	0,17	2,17	0,17	2,17	0,17
G	1,17	2,17	**3,17**	0,17	0,17	0,17	**8,17**	0,17	**5,17**	**4,17**	**5,17**	2,17	0,17
T	4,33	1,33	0,33	0,33	0,33	0,33	0,33	**8,33**	**3,33**	2,33	2,33	2,33	0,33
Sum	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	1,00

$$f'_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{i=1}^{A} n_{i,j} + k}$$

A *alphabet size (=4)*

$n_{i,j,}$ *occurrences of residue i at position j*

p_i *prior residue probability for residue i*

$f_{i,j}$ *relative frequency of residue i at position j*

k *pseudo weight (arbitrary, 1 in this case)*

$f'_{i,j}$ *corrected frequency of residue i at position j*

Frequencies corrected by a pseudo-count

Frequency matrix corrected with pseudo-count **k= 1** **Specific nucleotide frequencies**

Residue\position	1	2	3	4	5	6	7	8	9	10	11	12	Prior (p_i)
A	**0,148**	0,370	0,259	0,037	0,926	0,037	**0,037**	**0,037**	**0,037**	0,037	0,148	0,259	0,33
C	0,241	0,241	0,352	**0,908**	0,019	0,908	0,019	0,019	0,019	0,241	**0,019**	0,241	0,17
G	0,130	0,241	**0,352**	0,019	**0,019**	0,019	0,908	0,019	0,574	**0,463**	0,574	0,241	0,17
T	0,481	**0,148**	0,037	0,037	0,037	**0,037**	0,037	0,926	0,370	0,259	0,259	**0,259**	0,33
Sum	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,00

$$f'_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{i=1}^{A} n_{i,j} + k}$$

A	*alphabet size (=4)*
$n_{i,j,}$	*occurrences of residue i at position j*
p_i	*prior residue probability for residue i*
$f_{i,j}$	*relative frequency of residue i at position j*
k	*pseudo weight (arbitrary, 1 in this case)*
$f'_{i,j}$	*corrected frequency of residue i at position j*

Probability of a sequence segment under the matrix model

Scoring a sequence with a corrected frequency matrix **k= 1** **Specific nucleotide frequencies**

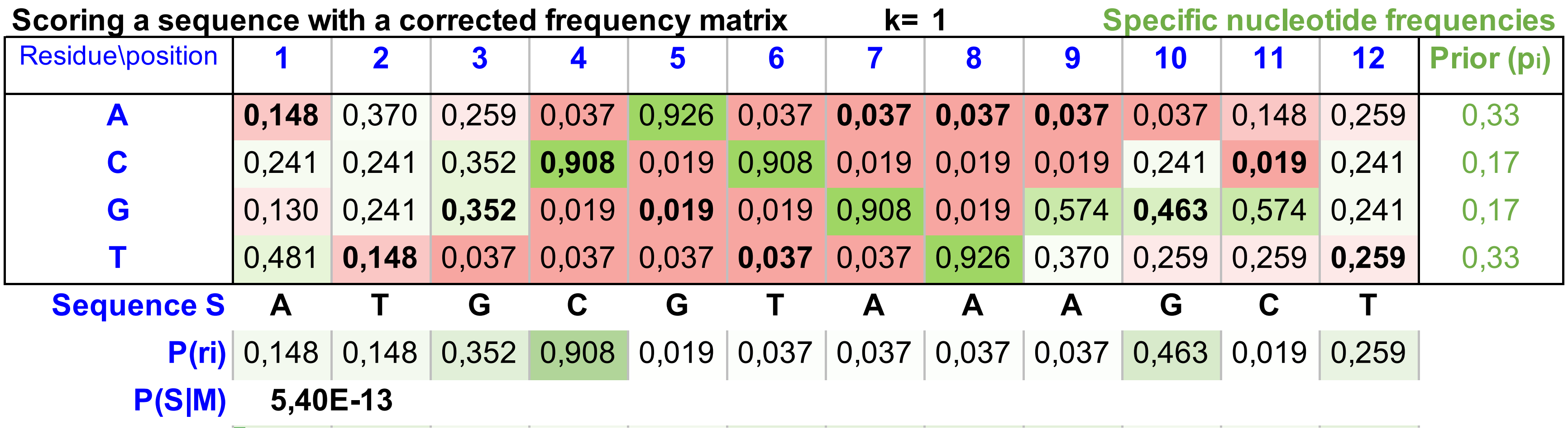
Residue\position	1	2	3	4	5	6	7	8	9	10	11	12	Prior (p_i)
A	**0,148**	0,370	0,259	0,037	0,926	0,037	**0,037**	**0,037**	**0,037**	0,037	0,148	0,259	0,33
C	0,241	0,241	0,352	**0,908**	0,019	0,908	0,019	0,019	0,019	0,241	**0,019**	0,241	0,17
G	0,130	0,241	**0,352**	0,019	**0,019**	0,019	0,908	0,019	0,574	**0,463**	0,574	0,241	0,17
T	0,481	**0,148**	0,037	0,037	0,037	**0,037**	0,037	0,926	0,370	0,259	0,259	**0,259**	0,33
Sequence S	**A**	**T**	**G**	**C**	**G**	**T**	**A**	**A**	**A**	**G**	**C**	**T**	
P(ri)													
P(S|M)													

Exercise: estimate the probability of sequence
ATGCGTAAAGCT
given the motif M

$$P(S \mid M) = \prod_{j=1}^{w} f'_{r_j j}$$

- Let
  - M be a frequency matrix of width w
  - $S = \{r_1, r_2, \ldots, r_w\}$ be a sequence segment of length w (same length as the matrix)
  - r_j is the residue found at position j of the sequence segment S.
- The corrected frequencies F'_{ij} can be used to estimate the probability to observe residue i at position j of the motif described by the matrix
- The probability to generate the sequence segment S under the model described by the matrix M is the product of the frequencies of residues at the corresponding columns of the matrix.

Probability of a sequence segment under the matrix model



Scoring a sequence with a corrected frequency matrix **k= 1** **Specific nucleotide frequencies**

Residue\position	1	2	3	4	5	6	7	8	9	10	11	12	Prior (pi)
A	**0,148**	0,370	0,259	0,037	0,926	0,037	**0,037**	**0,037**	**0,037**	0,037	0,148	0,259	0,33
C	0,241	0,241	0,352	**0,908**	0,019	0,908	0,019	0,019	0,019	0,241	**0,019**	0,241	0,17
G	0,130	0,241	**0,352**	0,019	**0,019**	0,019	0,908	0,019	0,574	**0,463**	0,574	0,241	0,17
T	0,481	**0,148**	0,037	0,037	0,037	**0,037**	0,037	0,926	0,370	0,259	0,259	**0,259**	0,33
Sequence S	**A**	**T**	**G**	**C**	**G**	**T**	**A**	**A**	**A**	**G**	**C**	**T**	
P(ri)	0,148	0,148	0,352	0,908	0,019	0,037	0,037	0,037	0,037	0,463	0,019	0,259	
P(S|M)	**5,40E-13**												

$$P(S \mid M) = \prod_{j=1}^{w} f'_{r_j j}$$

- Let
  - M be a frequency matrix of width w
  - $S = \{r_1, r_2, \ldots, r_w\}$ be a sequence segment of length w (same length as the matrix)
  - r_j is the residue found at position j of the sequence segment S.
- The corrected frequencies F'_{ij} can be used to estimate the probability to observe residue i at position j of the motif described by the matrix
- The probability to generate the sequence segment S under the model described by the matrix M is the product of the frequencies of residues at the corresponding columns of the matrix.

Background probability of a sequence segment – Bernoulli model

Residue	Prior (p_i)
A	0,33
C	0,17
G	0,17
T	0,33

$$P(S \mid B) = \prod_{j=1}^{w} p_{r_j}$$

Sequence S	A	T	G	C	G	T	A	A	A	G	C	T
p_i	0,330	0,330	0,170	0,170	0,170	0,330	0,330	0,330	0,330	0,170	0,170	0,330
P(S|B)	**6,05E-08**											

- A background model (B) should be defined to estimate the probability of a sequence motif outside of the motif.
- Various possibilities can be envisaged to define the background model
  - **Identical and independent distribution (iid):** Bernoulli model with equiprobable residues (this should generally be avoided, because most biological sequences are biased towards some residues)
  - **Bernoulli model with residue-specific probabilities** (p_r)
  - **Markov models** (treat dependencies between successive nucleotides)
- Under a Bernoulli model, the probability of a sequence motif S is the probability of the prior frequencies of its residues r_j.

The weight score

Scoring a sequence with a corrected frequency matrix **k= 1** **Specific nucleotide frequencies**

Residue\position	1	2	3	4	5	6	7	8	9	10	11	12	Prior (pi)
A	**0,148**	0,370	0,259	0,037	0,926	0,037	**0,037**	**0,037**	**0,037**	0,037	0,148	0,259	0,33
C	0,241	0,241	0,352	**0,908**	0,019	0,908	0,019	0,019	0,019	0,241	**0,019**	0,241	0,17
G	0,130	0,241	**0,352**	0,019	**0,019**	0,019	0,908	0,019	0,574	**0,463**	0,574	0,241	0,17
T	0,481	**0,148**	0,037	0,037	0,037	**0,037**	0,037	0,926	0,370	0,259	0,259	**0,259**	0,33
Sequence S	**A**	**T**	**G**	**C**	**G**	**T**	**A**	**A**	**A**	**G**	**C**	**T**	
P(ri)	0,148	0,148	0,352	0,908	0,019	0,037	0,037	0,037	0,037	0,463	0,019	0,259	
P(S|M)	**5,40E-13**												
pi	0,330	0,330	0,170	0,170	0,170	0,330	0,330	0,330	0,330	0,170	0,170	0,330	
P(S|B)	**6,05E-08**												
P(S|M)/P(S|B)	**8,93E-06**												
W(S)	**-11,63**												

$$W_S = \ln\left(\frac{P(S \mid M)}{P(S \mid B)}\right)$$

- The weight score is the log-likelihood of the sequence between two alternative generative models
  - The motif
  - The background model

The weight score: equivalent computing (only if Bernoulli model)

Scoring a sequence with a corrected frequency matrix **k= 1** **Specific nucleotide frequencies**

Residue\position	1	2	3	4	5	6	7	8	9	10	11	12	Prior (p_i)
A	**0,148**	0,370	0,259	0,037	0,926	0,037	**0,037**	**0,037**	**0,037**	0,037	0,148	0,259	0,33
C	0,241	0,241	0,352	**0,908**	0,019	0,908	0,019	0,019	0,019	0,241	**0,019**	0,241	0,17
G	0,130	0,241	**0,352**	0,019	**0,019**	0,019	0,908	0,019	0,574	**0,463**	0,574	0,241	0,17
T	0,481	**0,148**	0,037	0,037	0,037	**0,037**	0,037	0,926	0,370	0,259	0,259	**0,259**	0,33
Sequence S	**A**	**T**	**G**	**C**	**G**	**T**	**A**	**A**	**A**	**G**	**C**	**T**	
P(ri)	0,148	0,148	0,352	0,908	0,019	0,037	0,037	0,037	0,037	0,463	0,019	0,259	
P(S|M)	**5,40E-13**												
pi	*0,330*	0,330	0,170	0,170	0,170	0,330	0,330	0,330	0,330	0,170	0,170	0,330	
P(S|B)	**6,05E-08**												
P(S|M)/P(S|B)	**8,93E-06**												
W(S)	**-11,63**												

$$W_S = \ln\left(\frac{P(S \mid M)}{P(S \mid B)}\right)$$

$$W_S = \sum ln\left(\frac{r_i}{p_i}\right)$$

Equivalent computation: sum the log-ratios of residue probabilties

	1	2	3	4	5	6	7	8	9	10	11	12
P(ri)/pi	0,448	0,448	2,072	5,340	0,111	0,111	0,111	0,111	0,111	2,725	0,111	0,785
ln(P(ri)/pi)	-0,803	-0,803	0,728	1,675	-2,197	-2,197	-2,197	-2,197	-2,197	1,003	-2,197	-0,243
W(S)	**-11,63**											

Position-weight matrix

Residue\position	1	2	3	4	5	6	7	8	9	10	11	12	Prior (p_i)
A	-0,80	0,11	-0,24	-2,20	1,03	-2,20	-2,20	-2,20	-2,20	-2,20	-0,80	-0,24	0,33
C	0,35	0,35	0,73	1,68	-2,20	1,68	-2,20	-2,20	-2,20	0,35	-2,20	0,35	0,17
G	-0,27	0,35	0,73	-2,20	-2,20	-2,20	1,68	-2,20	1,22	1,00	1,22	0,35	0,17
T	0,38	-0,80	-2,20	-2,20	-2,20	-2,20	-2,20	1,03	0,11	-0,24	-0,24	-0,24	0,33
Sum	-0,345	0,010	-0,983	-4,916	-5,560	-4,916	-4,916	-5,560	-3,062	-1,088	-2,026	0,214	1,00

$$W_{i,j} = \ln\left(\frac{f'_{i,j}}{p_i}\right)$$

$$f'_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{i=1}^{A} n_{i,j} + k} \qquad \sum_{i=1}^{A} f'_{i,j} = 1$$

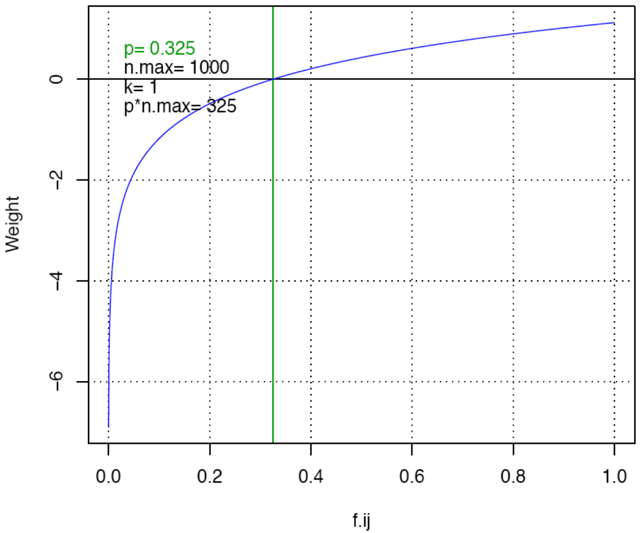
A alphabet size (=4)

p_i prior residue probability for residue i

$f_{i,j}$ relative frequency of residue i at position j

k pseudo weight (arbitrary, 1 in this case)

$f'_{i,j}$ corrected frequency of residue i at position j



Scoring a sequence with a weight matrix (Bernoulli only)

Residue\position	1	2	3	4	5	6	7	8	9	10	11	12	Prior (pi)
A	-0,80	0,11	-0,24	-2,20	1,03	-2,20	-2,20	-2,20	-2,20	-2,20	-0,80	-0,24	0,33
C	0,35	0,35	0,73	1,68	-2,20	1,68	-2,20	-2,20	-2,20	0,35	-2,20	0,35	0,17
G	-0,27	0,35	0,73	-2,20	-2,20	-2,20	1,68	-2,20	1,22	1,00	1,22	0,35	0,17
T	0,38	-0,80	-2,20	-2,20	-2,20	-2,20	-2,20	1,03	0,11	-0,24	-0,24	-0,24	0,33
Sum	-0,345	0,010	-0,983	-4,916	-5,560	-4,916	-4,916	-5,560	-3,062	-1,088	-2,026	0,214	1,00
Sequence S	A	T	G	C	G	T	A	A	A	G	C	T	
P(ri)	-0,80	-0,80	0,73	1,68	-2,20	-2,20	-2,20	-2,20	-2,20	1,00	-2,20	-0,24	
W(S)	-11,63												

$$W_S = \ln\left(\frac{P(S|M)}{P(S|B)}\right)$$

$$W_S = \ln\left(\frac{P(S|M)}{P(S|B)}\right) = \ln\left(\frac{\prod_{j=1}^{w} f'_{r_j j}}{\prod_{j=1}^{w} p_{r_j}}\right) = \sum_{j=1}^{w} \ln\left(\frac{f'_{r_j j}}{p_{r_j}}\right) = \sum_{j=1}^{w} W_{r_j j}$$

- The ***weight*** of a sequence segment is defined as the log-ratio between
  - *P(S|M)*, the sequence probability under the model described by the PSSM, and
  - *P(S|B),* the sequence probability under the background model.
- W_s represents the likelihood that segment *S* is an occurrence of the motif *M* rather than being issued from the background model *B*.
- Under Bernoulli assumption, the weight matrix W_{ij} can be used to simplify the computation of segment weights.

Ws	weight of sequence segment *S*
P(S|M)	probability of the sequence segment, given the matrix
P(S|B)	probability of the sequence segment, given the background
j	position within the segment and within the matrix
r_j	residue at position *j* of the sequence segment
p_{rj}	prior probability of residue r_j
f'_{rjj}	probability of residue r_j at position j of the matrix

Probability of the highest scoring sequence segment

Residue\position	1	2	3	4	5	6	7	8	9	10	11	12	Prior (p_i)
A	0,148	**0,370**	0,259	0,037	**0,926**	0,037	0,037	0,037	0,037	0,037	0,148	0,259	0,33
C	0,241	0,241	0,352	**0,908**	0,019	**0,908**	0,019	0,019	0,019	0,241	0,019	**0,241**	0,17
G	0,130	0,241	**0,352**	0,019	0,019	0,019	**0,908**	0,019	**0,574**	**0,463**	**0,574**	0,241	0,17
T	**0,481**	0,148	0,037	0,037	0,037	0,037	0,037	**0,926**	0,370	0,259	0,259	0,259	0,33

Sequence S	**T**	**C**	**C**	**C**	**A**	**C**	**G**	**T**	**G**	**G**	**G**	**C**
P(ri)	0,481	0,241	0,352	0,908	0,926	0,908	0,908	0,926	0,574	0,463	0,574	0,241
P(S|M)	**9,65E-04**											
pi	0,330	0,170	0,170	0,170	0,330	0,170	0,170	0,330	0,170	0,170	0,170	0,170
P(S|B)	**4,26E-09**											
P(S|M)/P(S|B)	**2,26E+05**											
W(S)	**12,33**											

Equivalent computation: sum the log-ratios of residue probabilties

P(ri)/pi	1,458	1,418	2,072	5,340	2,805	5,340	5,340	2,805	3,379	2,725	3,379	1,418
ln(P(ri)/pi)	0,377	0,349	0,728	1,675	1,031	1,675	1,675	1,031	1,218	1,003	1,218	0,349
W(S)	**12,33**											

This sequence has the highest possible weight score. Each nucleotide of the sequence corresponds to the residue with the highest weight in the corresponding column of the matrix.

Scanning a sequence with a weight matrix

- The weight matrix is successively aligned to each position of the sequence, and the score is the sum of weights for the letters aligned at each position (Hertz & Stormo, 1999).

Example: matching at a given location of a sequence

```
.....GCTGCACGTGGCCC..
```

Weight matrix

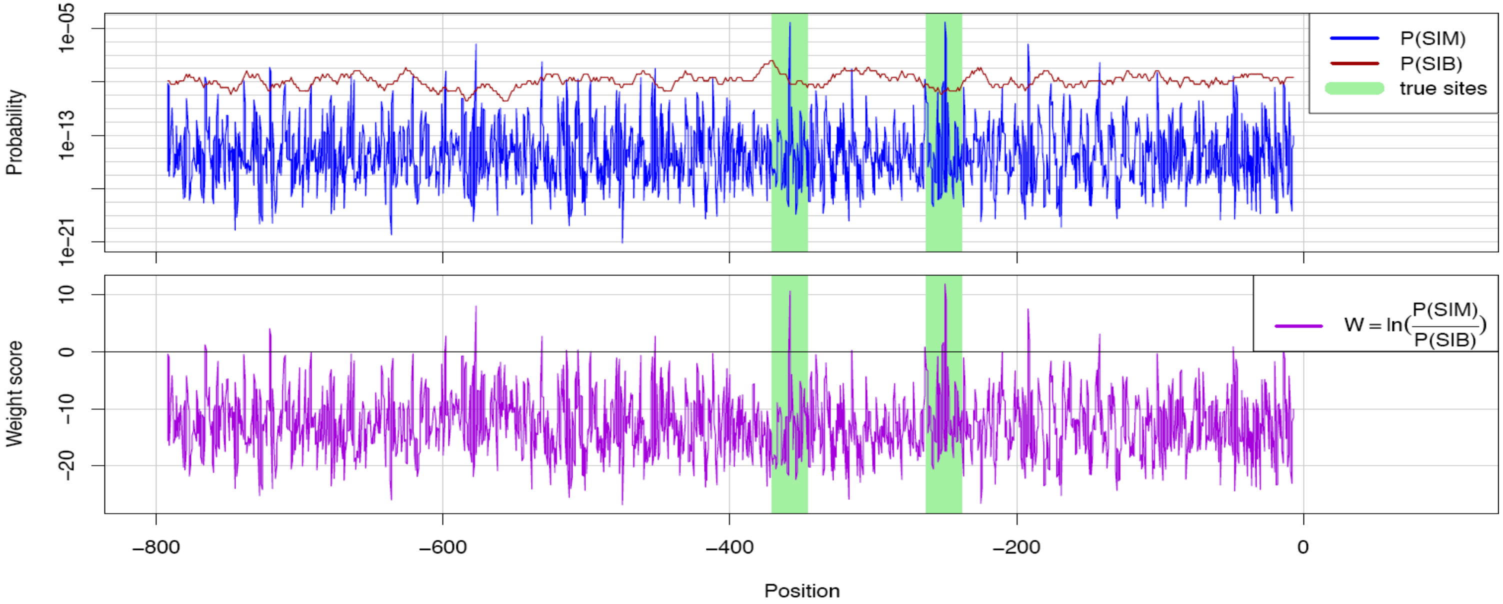
	1	2	3	4	5	6	7	8	9	10	11	12
A	-0.8	0.1	-0.2	-2.2	1.0	-2.2	-2.2	-2.2	-2.2	-2.2	-0.8	-0.2
C	0.3	0.3	0.7	1.6	-2.2	1.6	-2.2	-2.2	-2.2	0.3	-2.2	0.3
G	-0.3	0.3	0.7	-2.2	-2.2	-2.2	1.6	-2.2	1.2	1.0	1.2	0.3
T	0.4	-0.8	-2.2	-2.2	-2.2	-2.2	-2.2	1.0	0.1	-0.2	-0.2	-0.2

Scanning

1	**SUM**	G	C	T	G	C	A	C	G	T	G	G	C	C	C
	-10.54	-0.3	0.3	-2.2	-2.2	-2.2	-2.2	-2.2	-2.2	0.1	1.0	1.2	0.3		
2		C	T	G	C	A	C	G	T	G	G	C	C	C	
	7.55	0.3	-0.8	0.7	1.6	1.0	1.6	1.6	1.0	1.2	1.0	-2.2	0.3		

Scanning a sequence with a position-specific scoring matrix

- **P(S|M)** probability for site S to be generated as an instance of the motif.
- **P(S|B)** probability for site S to be generated as an instance of the background.
- **W** weight, i.e. the log ratio of the two above probabilities.
  - A positive weight indicates that a site is more likely to be an instance of the motif than of the background.



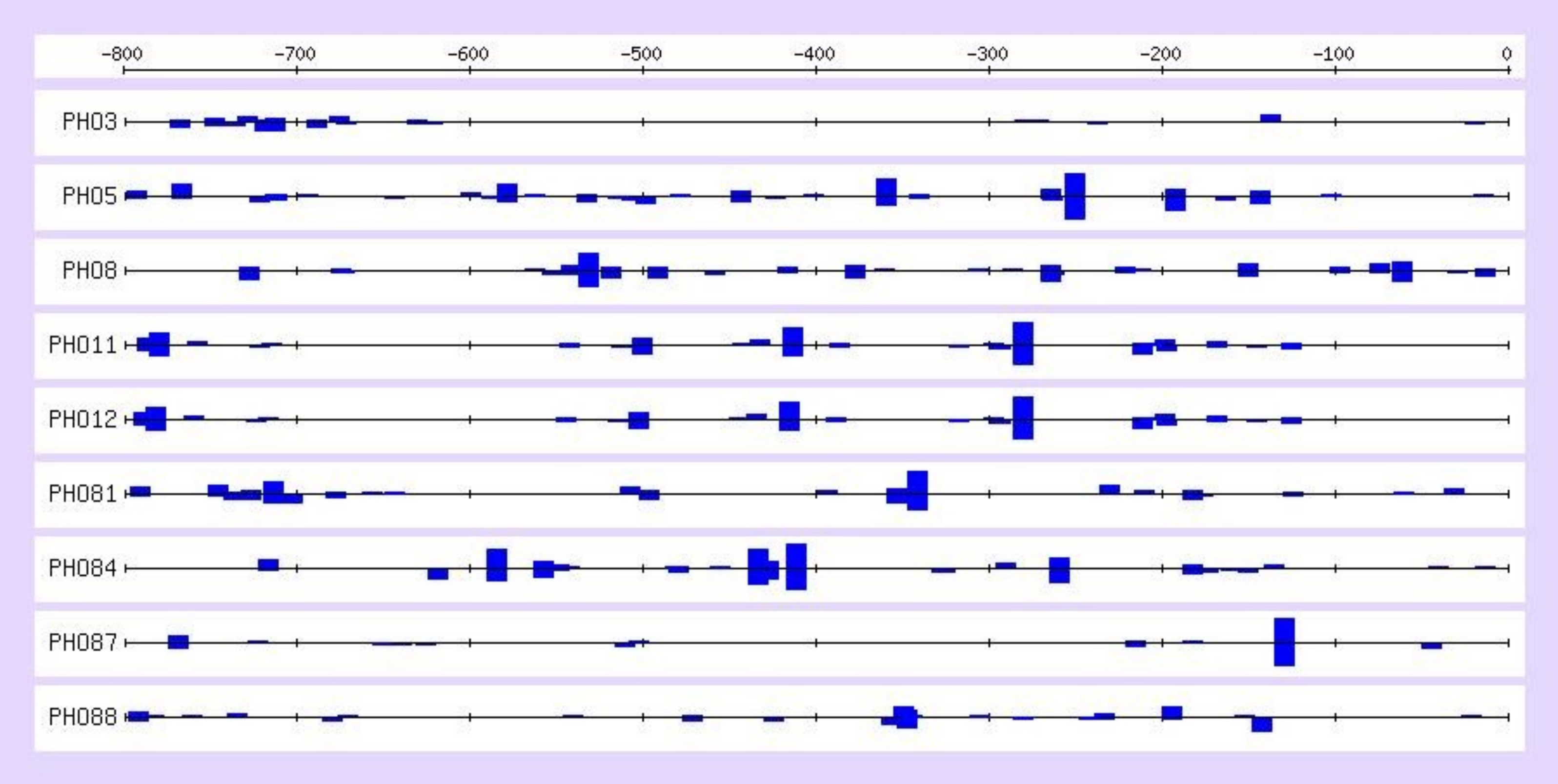
$$P(S \mid M) = \prod_{j=1}^{w} f'_{r_j j}$$

$$P(S \mid B) = \prod_{j=1}^{w} p_{r_j}$$

$$W_S = \ln\left(\frac{P(S \mid M)}{P(S \mid B)}\right)$$

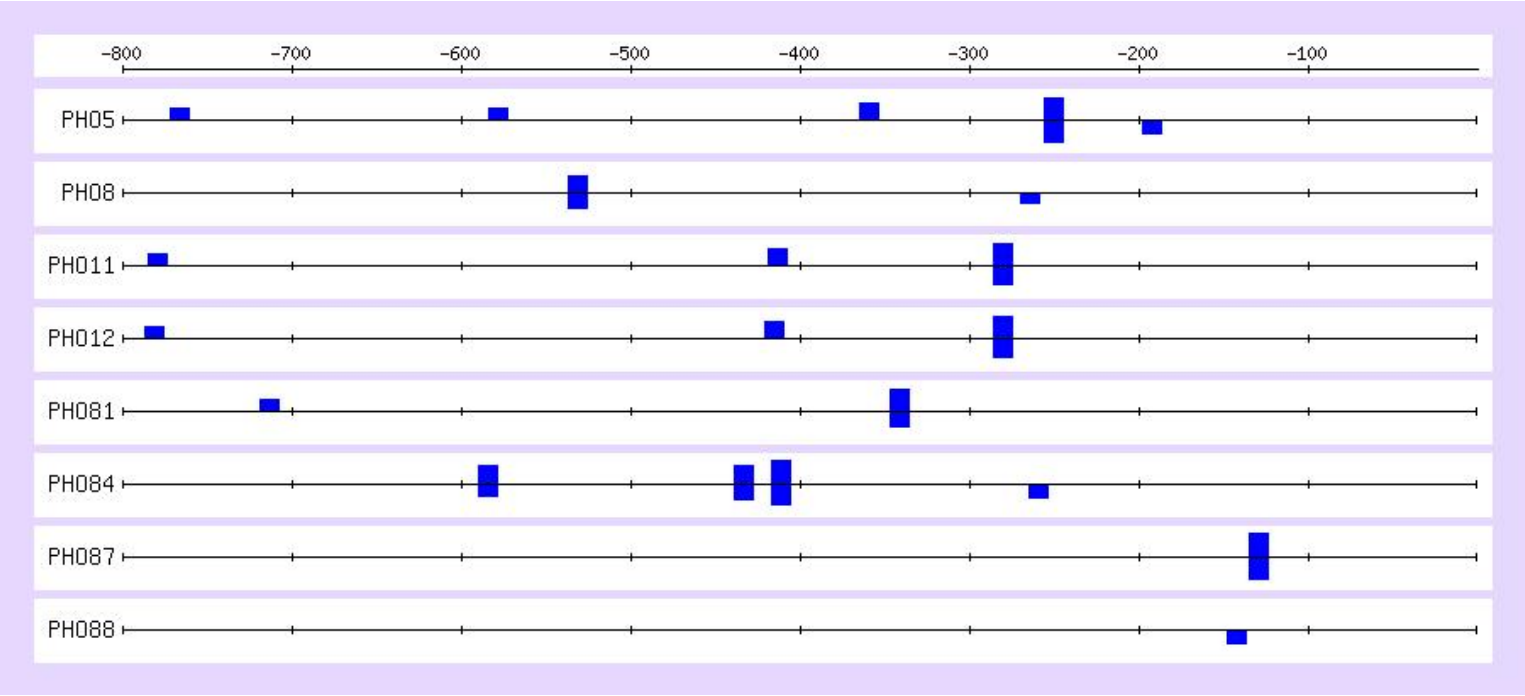
Matrix search : matching positions

- Matrix-based pattern matching is more sensitive than string-based pattern matching.
- How to choose the threshold ?



Matrix search : threshold choice

- The program Patser (G. Hertz) includes an option to automatically select a threshold on the basis of
  - the information content of the matrix
  - the length of the sequence to be scanned
- Another approach is to select the threshold on the basis of scores returned when the matrix is used to scan known binding sites for the factor.



Markov chains and transition matrices

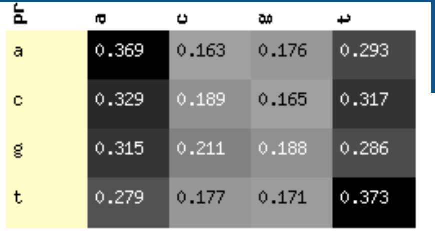
Transition matrix, order 1

Prefix/Suffi	A	C	G	T	N(Suffix)
a	0.369	0.163	0.176	0.293	**0.323**
c	0.329	0.189	0.165	0.317	**0.181**
g	0.315	0.211	0.188	0.286	**0.174**
t	0.279	0.177	0.171	0.373	**0.322**
; P_res	**0.323**	**0.181**	**0.174**	**0.322**	

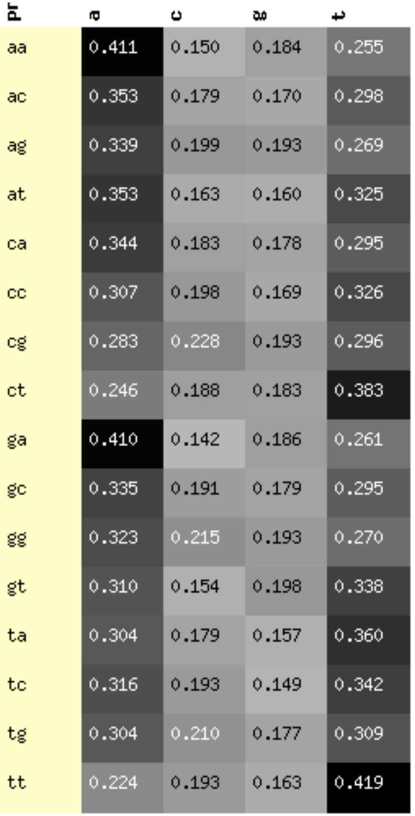
$$P\left(r_i \mid S_{i-m,i-1}\right)$$

Transition matrix, order 2

Prefix/Suffix	A	C	G	T	P(Prefix
aa	0.411	0.150	0.184	0.255	**0.119**
ac	0.353	0.179	0.170	0.298	**0.053**
ag	0.339	0.199	0.193	0.269	**0.057**
at	0.353	0.163	0.160	0.325	**0.095**
ca	0.344	0.183	0.178	0.295	**0.059**
cc	0.307	0.198	0.169	0.326	**0.034**
cg	0.283	0.228	0.193	0.296	**0.030**
ct	0.246	0.188	0.183	0.383	**0.057**
ga	0.410	0.142	0.186	0.261	**0.055**
gc	0.335	0.191	0.179	0.295	**0.037**
gg	0.323	0.215	0.193	0.270	**0.033**
gt	0.310	0.154	0.198	0.338	**0.050**
ta	0.304	0.179	0.157	0.360	**0.090**
tc	0.316	0.193	0.149	0.342	**0.057**
tg	0.304	0.210	0.177	0.309	**0.055**
tt	0.224	0.193	0.163	0.419	**0.120**
P(Suffix)	**0.323**	**0.181**	**0.174**	**0.322**	



pr	a	c	g	t
a	0.369	0.163	0.176	0.293
c	0.329	0.189	0.165	0.317
g	0.315	0.211	0.188	0.286
t	0.279	0.177	0.171	0.373



pr	a	c	g	t
aa	0.411	0.150	0.184	0.255
ac	0.353	0.179	0.170	0.298
ag	0.339	0.199	0.193	0.269
at	0.353	0.163	0.160	0.325
ca	0.344	0.183	0.178	0.295
cc	0.307	0.198	0.169	0.326
cg	0.283	0.228	0.193	0.296
ct	0.246	0.188	0.183	0.383
ga	0.410	0.142	0.186	0.261
gc	0.335	0.191	0.179	0.295
gg	0.323	0.215	0.193	0.270
gt	0.310	0.154	0.198	0.338
ta	0.304	0.179	0.157	0.360
tc	0.316	0.193	0.149	0.342
tg	0.304	0.210	0.177	0.309
tt	0.224	0.193	0.163	0.419

Scoring a sequence segment with a Markov model

- The example below illustrates the computation of the probability of a short seqence (ATGCGTAAAGCT) with a Markov chain of order 2, estimated from 3nt frequencies on the yeast genome.

The picture can’t be displayed.

The picture can’t be displayed.

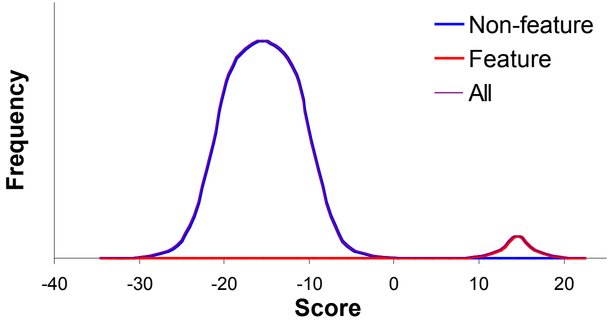
pos	P(R|W)		wR	S	P(S)
1	P(at)	0.094	at	at	9.42E-02
3	P(g|at)	0.161	atG	atg	1.52E-02
4	P(c|tg)	0.210	tgC	atgc	3.19E-03
5	P(g|gc)	0.180	gcG	atgcg	5.74E-04
6	P(t|cg)	0.295	cgT	atgcgt	1.69E-04
7	P(a|gt)	0.309	gtA	atgcgta	5.23E-05
8	P(a|ta)	0.304	taA	atgcgtaa	1.59E-05
9	P(a|aa)	0.409	aaA	atgcgtaaa	6.50E-06
10	P(g|aa)	0.184	aaG	atgcgtaaag	1.20E-06
11	P(c|ag)	0.200	agC	atgcgtaaagc	2.39E-07
12	P(t|gc)	0.294	gcT	atgcgtaaagct	7.04E-08

Sensitivity / selectivity tradeoff

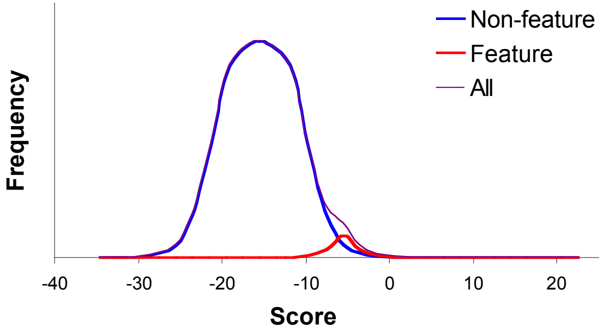
- The sequence is scanned with the matrix, and a score is assigned to each position.
- The highest score reflects the highest probability of having a functional site.
- How to define the threshold ? There is a tradeoff :
  - high selectivity ⇔ low sensitivity
  - high confidence in the predicted sites, but many real sites are missed
  - low selectivity ⇔ high sensitivity
    the real sites are drawn in a see of false positive

Discrimination power of a matrix

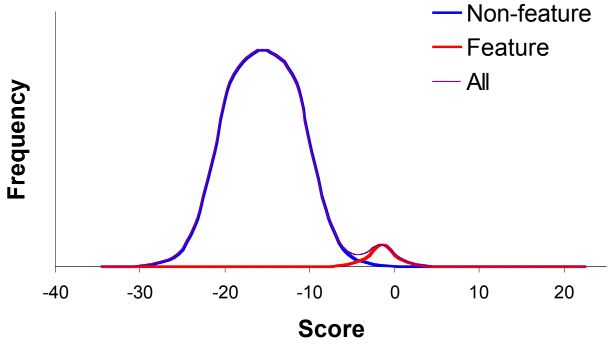
Highly discriminant



Poorly discriminant



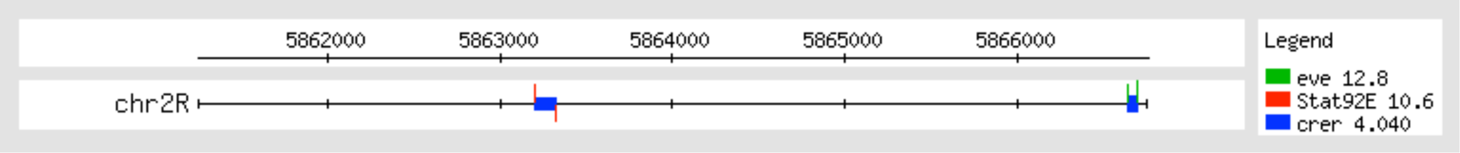
Reasonably discriminant



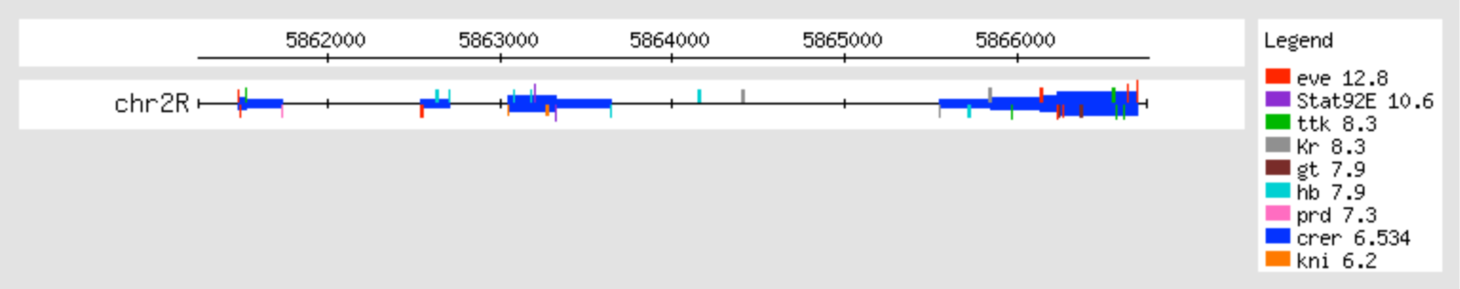
Exercise: impact of the P-value threshold on matrix scan results

- Open a connection to RSAT (http://www.rsat.eu/)
- Menu “Pattern matching”, tool “matrix scan (full options)”
- Click on DEMO2 to load the test case
  - Even skipped upstream sequences (5kb upstream of start codon)
  - Background model calibrated on Drosophila upstream sequences

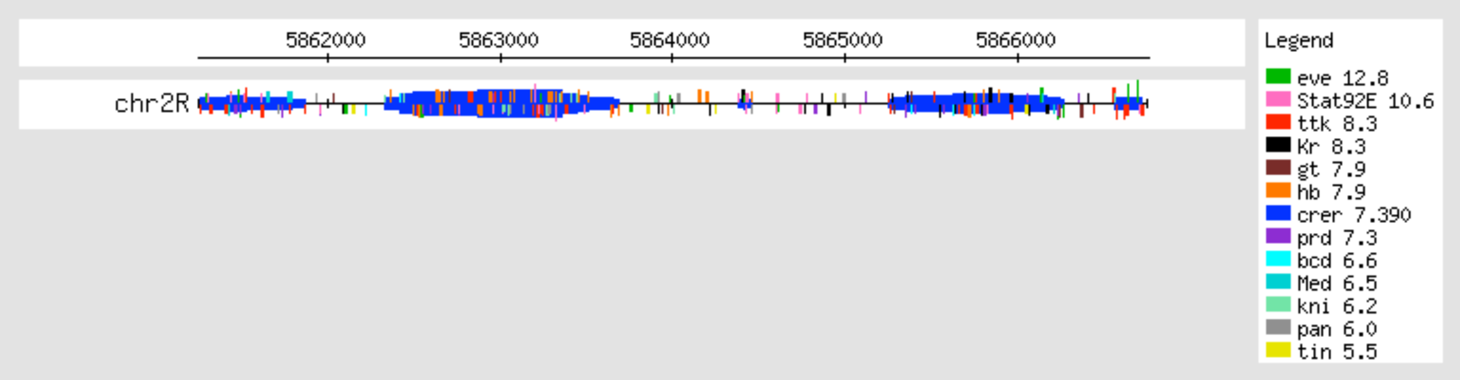
Impact of site P-value threshold on CRER detection



P-value <= 1e-3

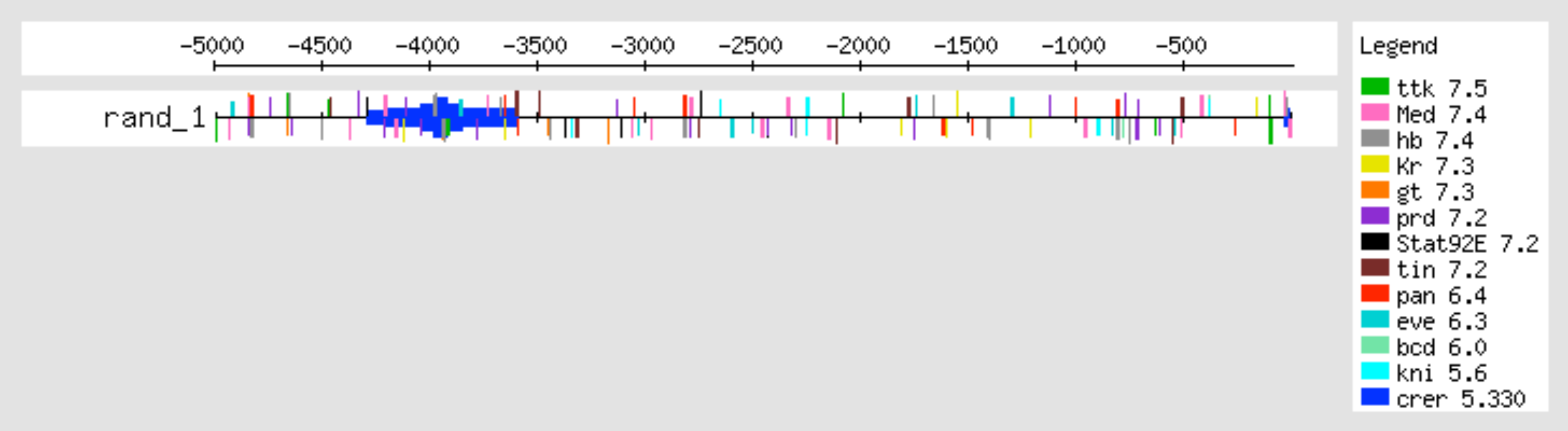


P-value <= 1e-4



P-value <= 1e-5

Negative control: CRERs in random sequences



Matching a sequence with a library of patterns

Match a sequence with a library of patterns

- Goal : given a sequence, find matches for any known regulatory site
  - → identify transcription factors that could regulate the gene
- Strategy: apply systematically pattern search with all patterns stored in the library
- Problem: how to set the threshold for the different patterns ?
- Warning : generates many false positive

Transfac Matsearch result - PHO5 upstream region

```
Inspecting sequence PHO5_4 [?] (1 - 816):
  F$NIT2_01           |      141 (+) | 1.000 |  0.995 | TATCtc
  F$PHO4_01           |      561 (+) | 1.000 |  0.990 | tcaCACGtggga
  F$PHO4_01           |      561 (-) | 1.000 |  0.982 | tccCACGtgtga
  F$NIT2_01           |      634 (+) | 1.000 |  0.972 | TATCaa
  F$NIT2_01           |      543 (-) | 1.000 |  0.967 | TATCga
  F$NIT2_01           |      676 (-) | 1.000 |  0.945 | TATCcc
  F$NIT2_01           |       31 (-) | 1.000 |  0.937 | TATCag
  F$PHO4_01           |      452 (+) | 1.000 |  0.935 | tagCACGttttc
  F$MCM1_01           |      666 (-) | 0.961 |  0.929 | tatCCCAaatgggtat
  F$MATA1_01          |      202 (+) | 1.000 |  0.926 | tGATGtcagt
  F$GCR1_01           |      323 (-) | 1.000 |  0.922 | gaCTTCcaa
  F$GCN4_C            |      536 (+) | 0.837 |  0.902 | aaaTGAATcg
  F$ABAA_01           |      292 (-) | 1.000 |  0.889 | atttgcgCATTcttgttga
  F$ABF_C             |      205 (+) | 0.887 |  0.885 | tgtcagtccccACGC
  F$MATA1_01          |      727 (+) | 1.000 |  0.882 | tGATGttttg
  F$MIG1_01           |      210 (-) | 1.000 |  0.881 | gctattagcgtGGGGac
  F$GCR1_01           |       69 (+) | 0.826 |  0.880 | ggCATCcaa
  F$PHO4_01           |       90 (-) | 1.000 |  0.879 | ggtCACGtttct
  F$MAT1MC_02         |      696 (+) | 1.000 |  0.875 | tgaaTTGTcg
  F$GCN4_C            |      589 (+) | 0.882 |  0.862 | ttaTGATTct
  F$STE11_01          |      415 (+) | 1.000 |  0.860 | ctttttCTTTgtctgcac
  F$GCR1_01           |      249 (-) | 0.783 |  0.859 | ggCGTCctg
  F$STE11_01          |      425 (-) | 1.000 |  0.859 | atatttCTTTgtgcagac
  F$MCM1_01           |      484 (+) | 0.831 |  0.855 | atgCCAAaaaaagtaa
```

Transfac Matsearch result - random sequence (mkv 5)

```
Inspecting sequence random mkv5 [?] (1 - 817):
  F$NIT2_01           |      176 (+) |  1.000 |  1.000 | TATCta
  F$NIT2_01           |      656 (+) |  1.000 |  1.000 | TATCta
  F$NIT2_01           |      275 (+) |  1.000 |  0.995 | TATCtc
  F$NIT2_01           |      455 (+) |  1.000 |  0.995 | TATCtc
  F$NIT2_01           |      298 (-) |  1.000 |  0.980 | TATCtt
  F$MATA1_01          |      506 (-) |  1.000 |  0.980 | tGATGtatgt
  F$ABF_C             |       84 (+) |  0.991 |  0.973 | aatcattcttgACGT
  F$MIG1_01           |      264 (-) |  1.000 |  0.958 | gagataaaactGGGGtt
  F$NIT2_01           |      701 (+) |  1.000 |  0.947 | TATCgt
  F$NIT2_01           |      802 (-) |  1.000 |  0.947 | TATCgt
  F$ABF1_01           |       81 (+) |  0.976 |  0.944 | gtaaatcattcttgACGTtttt
  F$MAT1MC_02         |      665 (-) |  1.000 |  0.918 | cctaTTGTga
  F$NIT2_01           |      280 (-) |  1.000 |  0.915 | TATCcg
  F$ABAA_01           |       42 (+) |  1.000 |  0.902 | tccccatCATTctaacagt
  F$PACC_01           |      331 (-) |  1.000 |  0.897 | acgaGCCAagaaagtt
  F$ABAA_01           |      201 (+) |  1.000 |  0.883 | accatagCATTctggatct
  F$MAT1MC_02         |      442 (-) |  1.000 |  0.882 | tataTTGTat
  F$ABF_C             |      638 (-) |  0.991 |  0.882 | agtcaaatgaaACGT
  F$ABF_C             |      609 (-) |  0.949 |  0.874 | tttctttaaacACGG
  F$MATA1_01          |      558 (-) |  1.000 |  0.868 | tGATGgaaga
  F$HSF_03            |      713 (-) |  1.000 |  0.859 | AGAAttgaaatttt
  F$MAT1MC_02         |      134 (-) |  1.000 |  0.858 | cacaTTGTgt
  F$ABAA_01           |       80 (+) |  1.000 |  0.856 | agtaaatCATTcttgacgt
  F$HAP234_01         |      332 (-) |  1.000 |  0.851 | acgagCCAAgaaaagt
```

Transfac Matsearch result - random sequence (iid)

```
Inspecting sequence random iid [?] (1 - 817):
  F$NIT2_01           |      534 (-) |  1.000 |  1.000 | TATCta
  F$NIT2_01           |      294 (+) |  1.000 |  0.995 | TATCtc
  F$NIT2_01           |      634 (-) |  1.000 |  0.972 | TATCaa
  F$NIT2_01           |      216 (-) |  1.000 |  0.965 | TATCtg
  F$STUAP_01          |      808 (-) |  1.000 |  0.959 | attCGCGtct
  F$NIT2_01           |       24 (+) |  1.000 |  0.952 | TATCat
  F$NIT2_01           |      343 (+) |  1.000 |  0.952 | TATCat
  F$NIT2_01           |      413 (-) |  1.000 |  0.952 | TATCat
  F$STUAP_01          |      441 (+) |  1.000 |  0.930 | aagCGCGcct
  F$NIT2_01           |      244 (-) |  1.000 |  0.930 | TATCct
  F$STUAP_01          |      808 (+) |  1.000 |  0.926 | agaCGCGaat
  F$GCR1_01           |      499 (+) |  1.000 |  0.922 | gaCTTCcta
  F$PACC_01           |      647 (-) |  1.000 |  0.920 | ctccGCCAggcactgaa
  F$NIT2_01           |      475 (+) |  1.000 |  0.915 | TATCcg
  F$ABF_C             |      235 (-) |  0.949 |  0.904 | tatcctgcaacACGG
  F$PHO4_01           |      246 (-) |  1.000 |  0.882 | gctCACGttatc
  F$GCR1_01           |      763 (-) |  1.000 |  0.866 | acCTTCcgc
  F$STUAP_01          |      441 (-) |  1.000 |  0.859 | aggCGCGctt
  F$MIG1_01           |      371 (+) |  1.000 |  0.857 | accgaaacagtGGGGtt
  F$MAT1MC_02         |      375 (-) |  0.769 |  0.855 | cccaCTGTtt
```