Regulatory sequence analysis

Matrix-based pattern matching

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Regulatory sites : matrix description

Alignment matrix

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		

Binding site for the yeast Pho4p transcription factor

(Source : Transfac matrix F$PHO4_01)

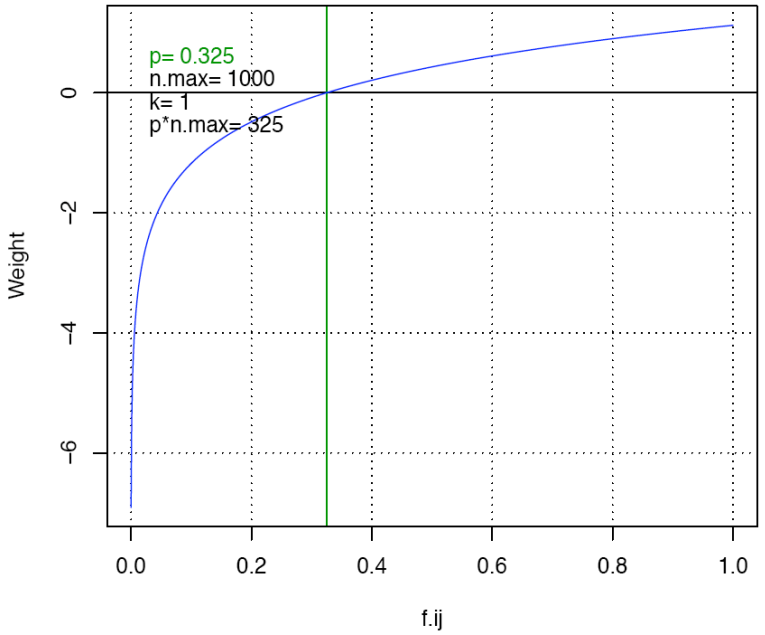
Position-weight matrix

Prior	Pos	1	2	3	4	5	6	7	8	9	10	11	12
0.33	A	-0.79	0.13	-0.23	-2.20	**1.05**	-2.20	-2.20	-2.20	-2.20	-2.20	-0.79	-0.23
0.18	C	0.32	0.32	**0.70**	**1.65**	-2.20	**1.65**	-2.20	-2.20	-2.20	0.32	-2.20	0.32
0.18	G	-0.29	0.32	**0.70**	-2.20	-2.20	-2.20	**1.65**	-2.20	**1.19**	**0.97**	**1.19**	0.32
0.33	T	0.39	-0.79	-2.20	-2.20	-2.20	-2.20	-2.20	**1.05**	0.13	-0.23	-0.23	-0.23
1	Sum	-0.37	-0.02	-1.02	-4.94	-5.55	-4.94	-4.94	-5.55	-3.08	-1.13	-2.03	0.19

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

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

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



Information content

Prior	Pos	1	2	3	4	5	6	7	8	9	10	11	12
0.33	A	-0.12	0.05	-0.06	-0.08	**0.97**	-0.08	-0.08	-0.08	-0.08	-0.08	-0.12	-0.06
0.18	C	0.08	0.08	**0.25**	**1.50**	-0.04	**1.50**	-0.04	-0.04	-0.04	0.08	-0.04	0.08
0.18	G	-0.04	0.08	**0.25**	-0.04	-0.04	-0.04	**1.50**	-0.04	**0.68**	**0.45**	**0.68**	0.08
0.33	T	0.19	-0.12	-0.08	-0.08	-0.08	-0.08	-0.08	**0.97**	0.05	-0.06	-0.06	-0.06
1	Sum	0.11	0.09	**0.36**	**1.29**	**0.80**	**1.29**	**1.29**	**0.80**	**0.61**	**0.39**	**0.47**	0.04

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

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

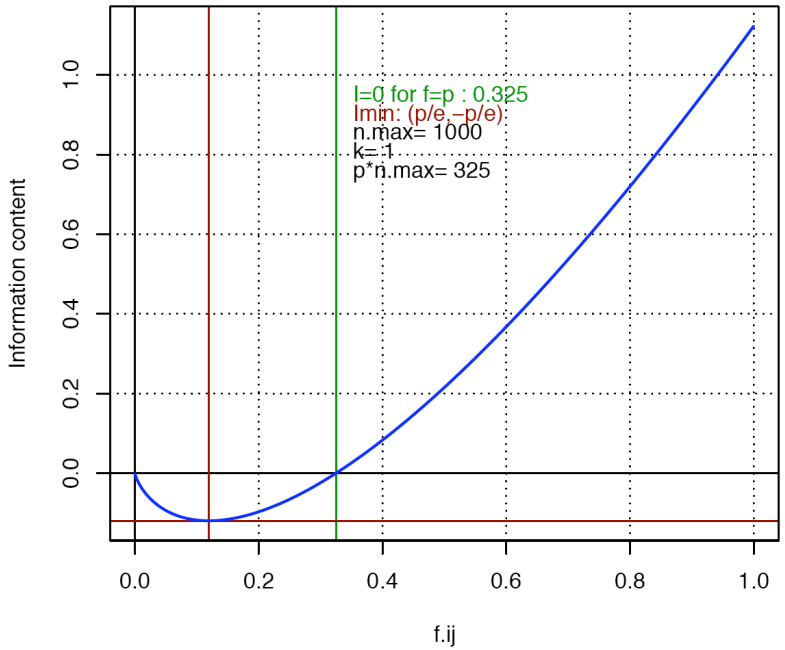
$$I_j = \sum_{i=1}^{A} I_{i,j}$$

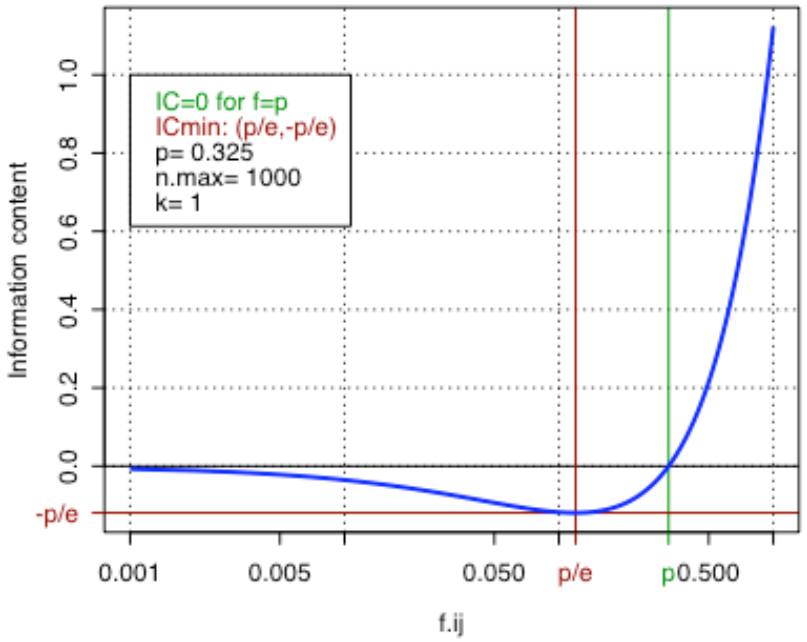
$$I_{matrix} = \sum_{j=1}^{w} \sum_{i=1}^{A} I_{i,j}$$

A	*alphabet size (=4 for DNA)*
w	*matrix width (=12)*
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*

Information content I_{ij} of a cell of the matrix

- I_{ij} is positive when $f'_{ij} > p_i$
  (i.e. when residue *i* is more frequent at position *j* than expected by chance)
- I_{ij} is negative when $f'_{ij} < p_i$
- I_{ij} tends towards 0 when f'_{ij} -> 0 (because $limit_{x->0}\ x*ln(x) = 0$)





Information content of a column of the matrix

- I_j is always positive
- I_j is 0 when the frequency of all residues equal their prior probability ($f_{ij}=p_i$)
- I_j is maximal when
  - the residue i_m with the lowest prior probability has a frequency of 1 (all other residues have a frequency of 0)
  - and the pseudo-weight is 0

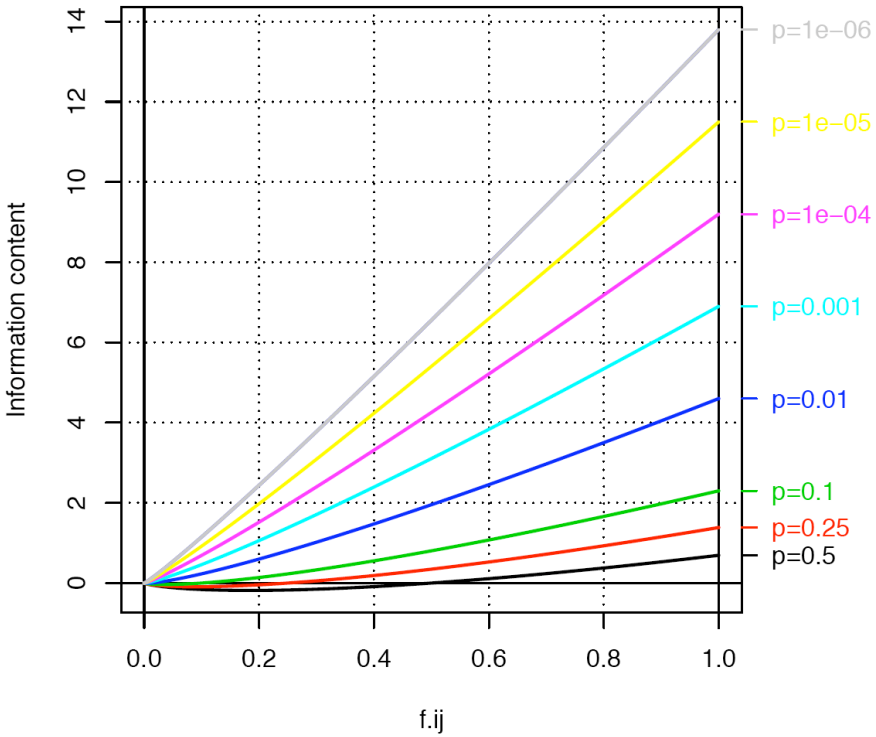
$$I_j = \sum_{i=1}^{A} I_{i,j} = \sum_{i=1}^{A} f'_{i,j} \ln\left(\frac{f'_{i,j}}{p_i}\right)$$

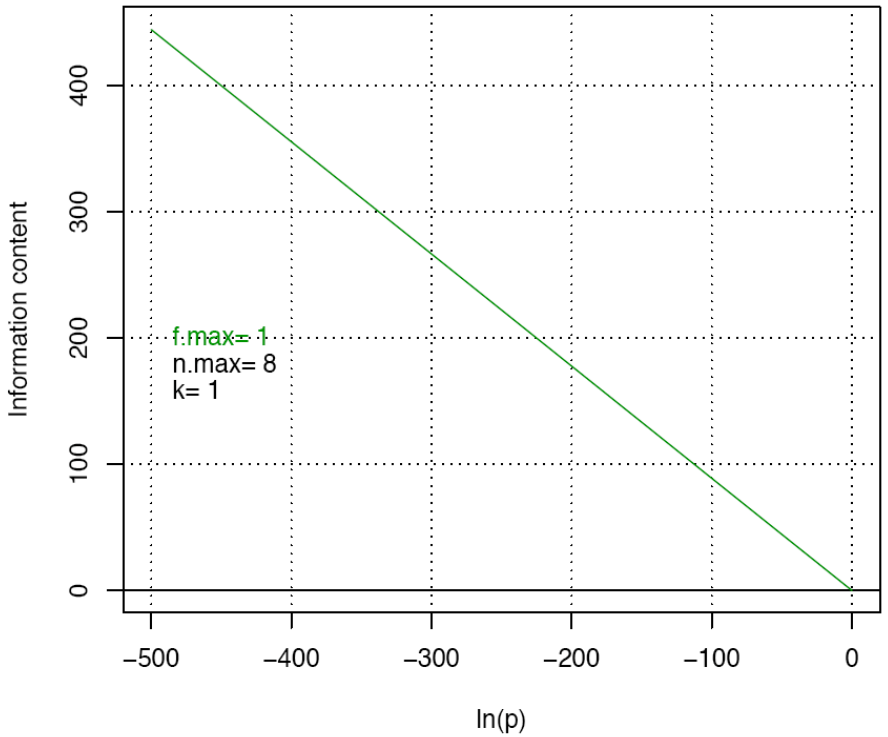
$$i_m = \arg\min_i(p_i) \qquad k = 0$$

$$\max(I_j) = 1 * \ln(\frac{1}{p_i}) = -\ln(p_i)$$

Information content: effect of prior probabilities

- The upper bound of I_j increases when p_i decreases
  - $I_j \text{ -> } Inf$ when $p_i \text{ -> } 0$
- The information content, as defined by Gerald Hertz, has thus no upper bound.





References - PSSM information content

- Papers by Tom Schneider
  - Schneider, T.D., G.D. Stormo, L. Gold, and A. Ehrenfeucht. 1986. Information content of binding sites on nucleotide sequences. *J Mol Biol* **188:** 415-431.
  - Tom Schneider's publications online
    - http://www.lecb.ncifcrf.gov/~toms/paper/index.html
- Papers by Gerald Hertz
  - Hertz, G.Z. and G.D. Stormo. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15:** 563-577.

Scanning a sequence with a profile 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).

Ex: 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**		
3		**T**	**G**	**C**	**A**	**C**	**G**	**T**	**G**	**G**	**C**	**C**	**C**		
	-9.93	**0.4**	**0.3**	**0.7**	**-2.2**	**-2.2**	**-2.2**	**-2.2**	**-2.2**	**1.2**	**0.3**	**-2.2**	**0.3**		

Interpretation of the matching weight

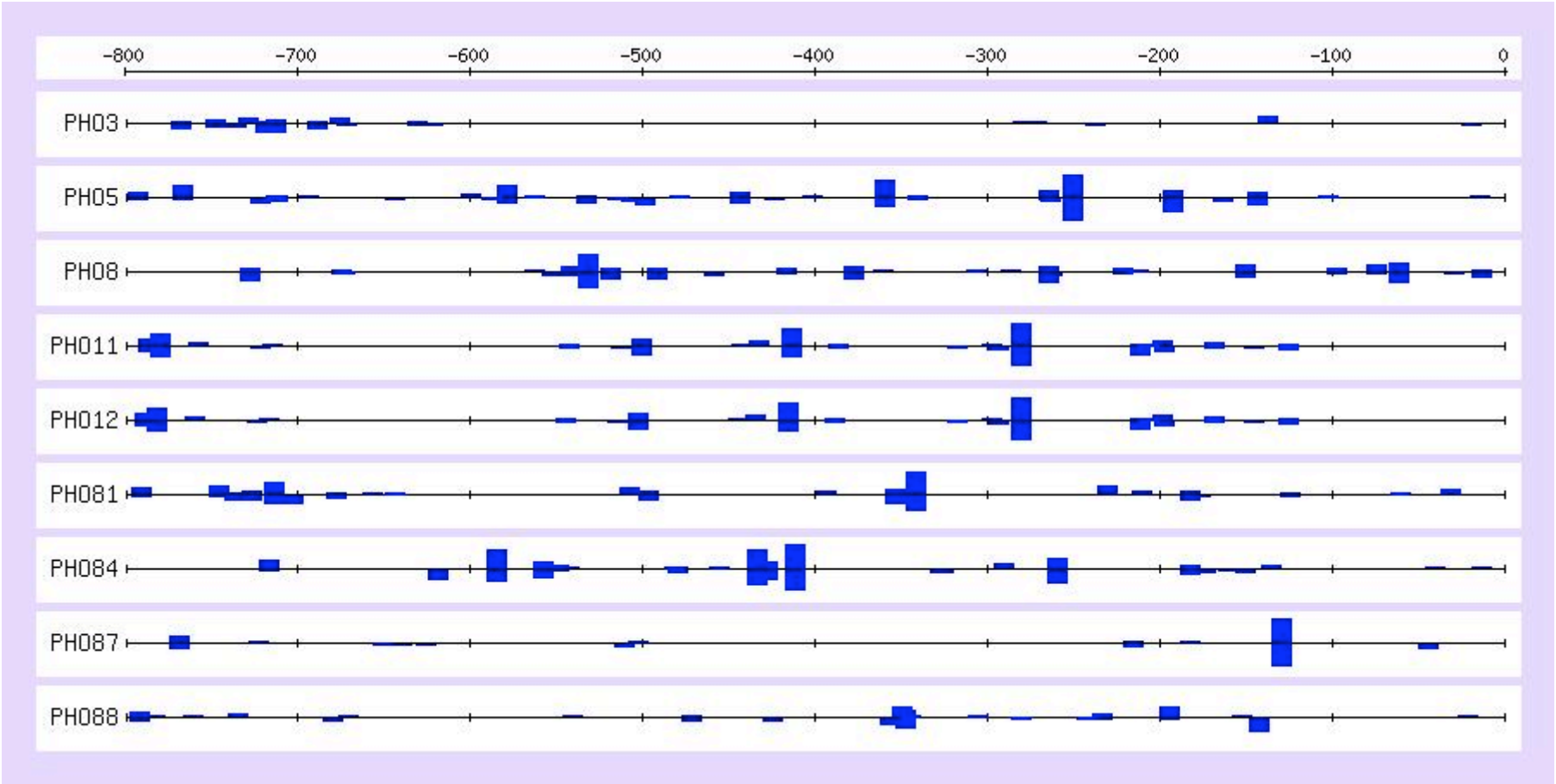
$$W_S = \sum_{k=1}^{w} W_{r_k k} = \sum_{k=1}^{w} \ln\left(\frac{f'_{r_k k}}{p_{r_k}}\right) = \ln\left(\prod_{k=1}^{w} \frac{f'_{r_k k}}{p_{r_k}}\right) = \ln\left(\frac{\prod_{k=1}^{w} f'_{r_k k}}{\prod_{k=1}^{w} p_{r_k}}\right) = \ln\left(\frac{P(S \mid M)}{P(S \mid B)}\right)$$

- The matching between a matrix and a segment of sequence is the sum of weights of the aligned residues.
- This is equivalent to the logarithm of the ratio between
  - (1) the product of the matrix frequencies, and
  - (2) the product of the prior probabilities for the residues found in the sequence segment.
- The term (1) is the probability to observe the sequence segment within the motif described by the PSSM. The term (2) is the probability to observe the sequence segment within the background.
- In other terms, the segment weight is the log likelihood ratio of its probability to be found within/without the motif. It indicates the likelihood to be within a motif when we observe that sequence segment.

Ws	weight of sequence segment S
k	position within the alignment
r_k	residue at position k of the sequence segment
p_{rk}	prior probability of residue r_k
f'_{rkk}	probability of residue r_k at positio k of the matric
$P(S\|M)$	probability of the sequence segment, given the matrix
$P(S\|B)$	probability of the sequence segment, given the background

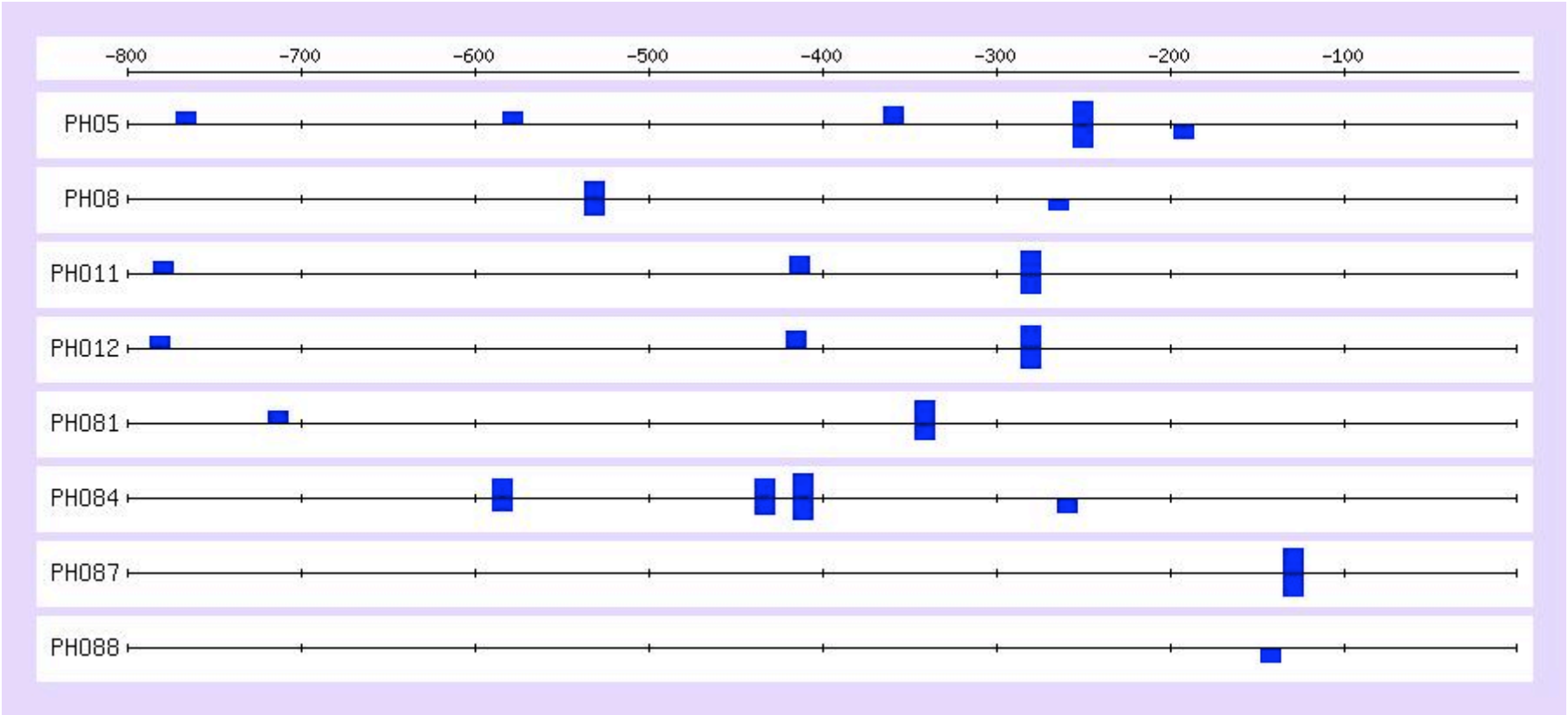
Matrix search : matching positions

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

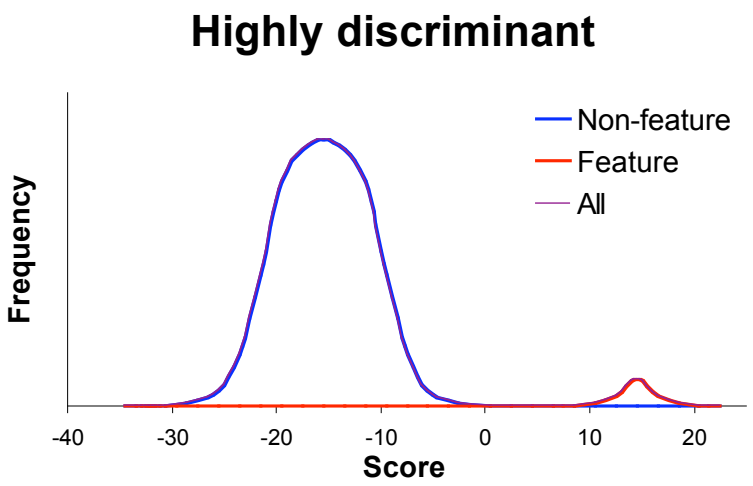


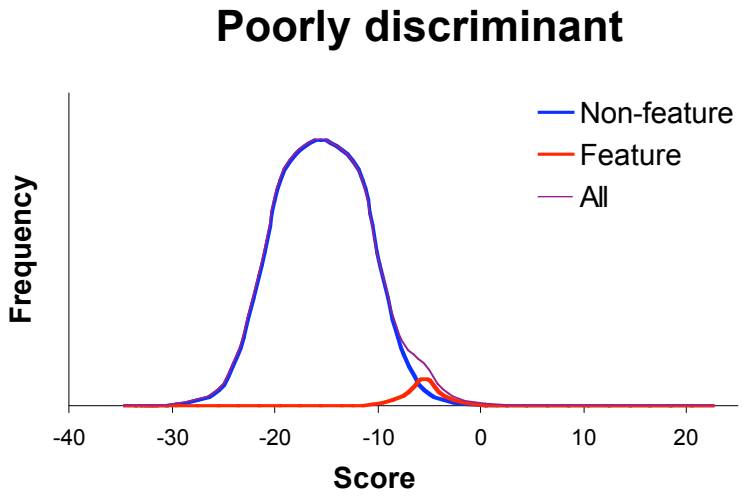
Matrix search : threshold selection

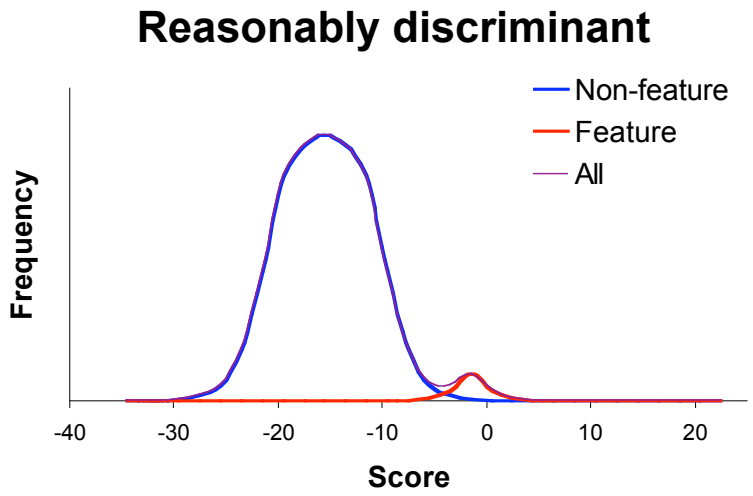
- Patser includes an option to automatically select a threshold on the basis of
  - the information content of the matrix
  - the length of the sequence to be scanned
- Note : the gene PHO3 is not displayed because there was not a single match. This gene is indeed not regulated by phosphate.
- Another approach would be to select the threshold on the basis of scores returned when the matrix is used to scan known binding sites for the factor.



Discrimination power of a matrix







Matrix search

- 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 trade :
  - 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