Regulatory sequence analysis

Matrix-based pattern matching

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

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

Pos	1	2	3	4	5	6	7	8	9	10	11	12
A	0.13	0.38	0.25	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00	0.13	0.25
C	0.25	0.25	**0.38**	**1.00**	0.00	**1.00**	0.00	0.00	0.00	0.25	0.00	0.25
G	0.13	0.25	**0.38**	0.00	0.00	0.00	**1.00**	0.00	**0.63**	**0.50**	**0.63**	0.25
T	0.50	0.13	0.00	0.00	0.00	0.00	0.00	**1.00**	0.38	0.25	0.25	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

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

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*

Reference: Hertz (1999). Bioinformatics 15:563-577.

Pseudo-count correction

Pos	1	2	3	4	5	6	7	8	9	10	11	12
A	0.15	0.37	0.26	0.04	**0.93**	0.04	0.04	0.04	0.04	0.04	0.15	0.26
C	0.24	0.24	**0.35**	**0.91**	0.02	**0.91**	0.02	0.02	0.02	0.24	0.02	0.24
G	0.13	0.24	**0.35**	0.02	0.02	0.02	**0.91**	0.02	**0.58**	**0.46**	**0.58**	0.24
T	0.48	0.15	0.04	0.04	0.04	0.04	0.04	**0.93**	0.37	0.26	0.26	0.26
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

1st option: identically distributed pseudo-weight

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

2nd option: pseudo-weight distributed according to residue priors

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

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*

Reference: Hertz (1999). Bioinformatics 15:563-577.

Probability of a sequence segment under the matrix model

Pos	1	2	3	4	5	6	7	8	9	10	11	12
A	**0.15**	0.37	0.26	0.04	0.93	0.04	**0.04**	**0.04**	**0.04**	0.04	0.15	0.26
C	0.24	0.24	0.35	**0.91**	0.02	0.91	0.02	0.02	0.02	0.24	**0.02**	0.24
G	0.13	0.24	**0.35**	0.02	**0.02**	0.02	0.91	0.02	0.58	**0.46**	0.58	0.24
T	0.48	**0.15**	0.04	0.04	0.04	**0.04**	0.04	0.93	0.37	0.26	0.26	**0.26**
Sequence S	**A**	**T**	**G**	**C**	**G**	**T**	**A**	**A**	**A**	**G**	**C**	**T**
P(res)	**0.15**	**0.15**	**0.35**	**0.91**	**0.02**	**0.04**	**0.04**	**0.04**	**0.04**	**0.46**	**0.02**	**0.26**
P(S|M)	**5.32E-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.

Probability of the best sequence segment under the matrix model

Pos	1	2	3	4	5	6	7	8	9	10	11	12
A	0.15	**0.37**	0.26	0.04	**0.93**	0.04	0.04	0.04	0.04	0.04	0.15	0.26
C	0.24	0.24	0.35	**0.91**	0.02	**0.91**	0.02	0.02	0.02	0.24	0.02	0.24
G	0.13	0.24	**0.35**	0.02	0.02	0.02	**0.91**	0.02	**0.58**	**0.46**	**0.58**	0.24
T	**0.48**	0.15	0.04	0.04	0.04	0.04	0.04	**0.93**	0.37	0.26	0.26	**0.26**
Sequence S	**T**	**A**	**G**	**C**	**A**	**C**	**G**	**T**	**G**	**G**	**G**	**T**
P(res)	**0.48**	**0.37**	**0.35**	**0.91**	**0.93**	**0.91**	**0.91**	**0.93**	**0.58**	**0.46**	**0.58**	**0.26**

P(S|M) **1.59E-03**

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

This segment of sequence is associated to the highest possible probability given the matrix : P(S|M)
Each nucleotide of the sequence corresponds to the residue with the highest probability in the corresponding column of the matrix.

Probability of a sequence segment under a Bernoulli background model

Pos	Prior
A	0.325
C	0.175
G	0.175
T	0.325

Sequence S	**A**	**T**	**G**	**C**	**G**	**T**	**A**	**A**	**A**	**G**	**C**	**T**
P(res)	**0.325**	**0.325**	**0.175**	**0.175**	**0.175**	**0.325**	**0.325**	**0.325**	**0.325**	**0.175**	**0.175**	**0.325**
P(S|B)	**6.29E-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
  - 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 chains
- Under a Bernoulli model, the probability of a sequence motif S is the probability of the prior frequencies of its residues r_j.

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

Weight of a sequence segment

Pos	1	2	3	4	5	6	7	8	9	10	11	12
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
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
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
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
residue r	A	T	G	C	G	T	A	A	A	G	C	T
W(r)	-0.79	-0.79	0.70	1.65	-2.20	-2.20	-2.20	-2.20	-2.20	0.97	-2.20	-0.23
Weight	-11.67		=SUM[W(r)]									

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

- 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.
- The weight 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 asumption, the weight matrix W_{ij} can be used to simplify the computation of segment weights.

Under assumption of a Bernoulli background model, this formula becomes

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

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

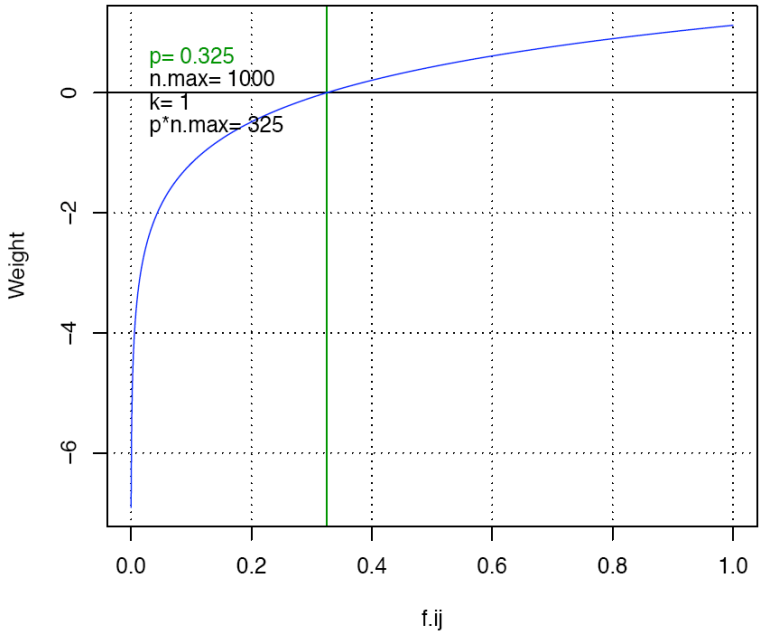
Position-weight matrix

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



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).

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		

Markov chains and transition matrices

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

- The two tables below show the transition matrices for a Markov model of order 1 (top) and 2 (bottom), respectively.
- The two models were trained with yeast non-coding upstream sequences.
- Notice the strong probability of transitions from AA to A and TT to T.

Transition matrix, order 1

Pre/Suffix	A	C	G	T	P(Prefix)	N(Suffix)
A	0.359	0.171	0.187	0.283	**0.313**	**1,467,035**
C	0.329	0.199	0.163	0.308	**0.191**	**894,658**
G	0.317	0.211	0.198	0.274	**0.187**	**874,683**
T	0.255	0.193	0.193	0.359	**0.310**	**1,451,623**
P(Suffix)	**0.313**	**0.191**	**0.187**	**0.310**		
N(Suffix)	**1,467,895**	**894,543**	**874,444**	**1,451,117**		

Transition matrix, order 2

Prefix/Suffix	A	C	G	T	P(Prefix	N(Prefix
AA	0.388	0.161	0.200	0.251	**0.112**	**525,000**
AC	0.339	0.198	0.173	0.290	**0.054**	**251,072**
AG	0.345	0.204	0.196	0.255	**0.059**	**274,601**
AT	0.311	0.184	0.182	0.323	**0.088**	**413,946**
CA	0.347	0.178	0.189	0.286	**0.063**	**293,750**
CC	0.341	0.190	0.161	0.309	**0.038**	**178,110**
CG	0.293	0.221	0.196	0.290	**0.031**	**145,876**
CT	0.229	0.195	0.205	0.371	**0.059**	**275,634**
GA	0.394	0.155	0.187	0.264	**0.059**	**277,053**
GC	0.330	0.205	0.169	0.297	**0.039**	**184,192**
GG	0.318	0.217	0.187	0.277	**0.037**	**173,266**
GT	0.285	0.175	0.204	0.336	**0.051**	**239,384**
TA	0.300	0.193	0.168	0.339	**0.079**	**369,426**
TC	0.313	0.203	0.152	0.332	**0.060**	**280,131**
TG	0.302	0.209	0.208	0.282	**0.060**	**279,783**
TT	0.210	0.208	0.189	0.392	**0.111**	**520,906**
P(Suffix)	**0.313**	**0.191**	**0.187**	**0.310**		
N(suffix)	**1,466,075**	**893,444**	**873,260**	**1,449,351**		

Scoring a sequence segment with a Markov model

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

$$P(S) = P(S_{1,m}) \prod_{i=m+1}^{L} P\left(r_i \mid S_{i-m,i-1}\right)$$

Transition matrix, order 2

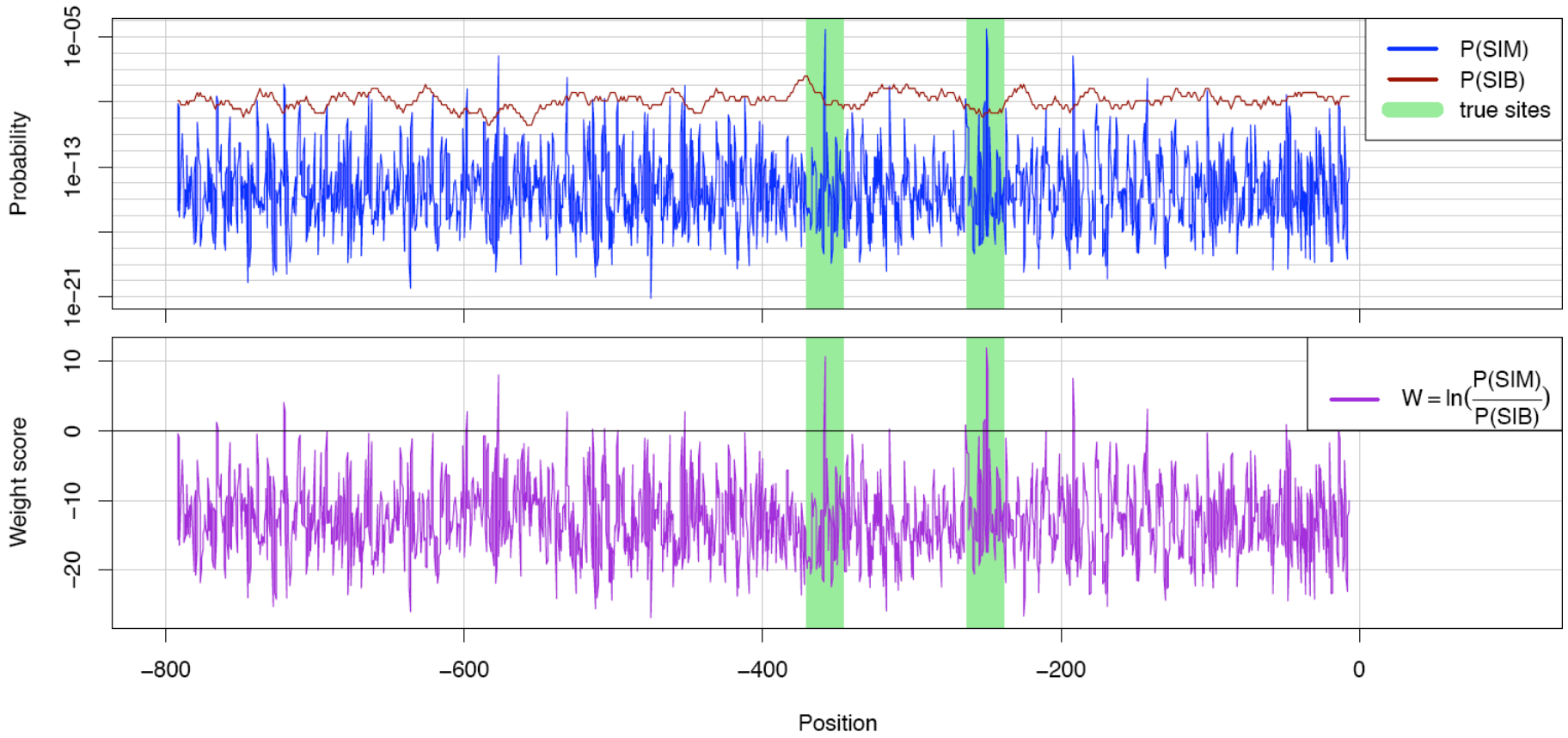
Prefix/Suffix	A	C	G	T	P(Prefix	N(Prefix
AA	0.388	0.161	0.200	0.251	**0.112**	**525,000**
AC	0.339	0.198	0.173	0.290	**0.054**	**251,072**
AG	0.345	0.204	0.196	0.255	**0.059**	**274,601**
AT	0.311	0.184	0.182	0.323	**0.088**	**413,946**
CA	0.347	0.178	0.189	0.286	**0.063**	**293,750**
CC	0.341	0.190	0.161	0.309	**0.038**	**178,110**
CG	0.293	0.221	0.196	0.290	**0.031**	**145,876**
CT	0.229	0.195	0.205	0.371	**0.059**	**275,634**
GA	0.394	0.155	0.187	0.264	**0.059**	**277,053**
GC	0.330	0.205	0.169	0.297	**0.039**	**184,192**
GG	0.318	0.217	0.187	0.277	**0.037**	**173,266**
GT	0.285	0.175	0.204	0.336	**0.051**	**239,384**
TA	0.300	0.193	0.168	0.339	**0.079**	**369,426**
TC	0.313	0.203	0.152	0.332	**0.060**	**280,131**
TG	0.302	0.209	0.208	0.282	**0.060**	**279,783**
TT	0.210	0.208	0.189	0.392	**0.111**	**520,906**
P(Suffix)	**0.313**	**0.191**	**0.187**	**0.310**		
N(suffix)	**1,466,075**	**893,444**	**873,260**	**1,449,351**		

pos	P(R|W)		wR	S	P(S)
1	P(CC)	0.038	cc	CC	3.80E-02
2	P(T|CC)	0.309	ccT	CCT	1.17E-02
3	P(A|CT)	0.229	ctA	CCTA	2.69E-03
4	P(C|TA)	0.193	taC	CCTAC	5.19E-04
5	P(T|AC)	0.290	acT	CCTACT	1.50E-04
6	P(A|CT)	0.229	ctA	CCTACTA	3.45E-05
7	P(T|TA)	0.339	taT	CCTACTAT	1.17E-05
8	P(A|AT)	0.311	atA	CCTACTATA	3.63E-06
9	P(T|TA)	0.339	taT	CCTACTATAT	1.23E-06
10	P(G|AT)	0.182	atG	CCTACTATATG	2.25E-07
11	P(C|TG)	0.209	tgC	CCTACTATATGC	4.69E-08
12	P(C|GC)	0.205	gcC	CCTACTATATGCC	9.61E-09
13	P(C|CC)	0.190	ccC	CCTACTATATGCCC	1.82E-09
14	P(A|CC)	0.341	ccA	CCTACTATATGCCCA	6.21E-10
15	P(G|CA)	0.189	caG	CCTACTATATGCCCAG	1.17E-10
16	P(A|AG)	0.345	agA	CCTACTATATGCCCAGA	4.04E-11
17	P(A|GA)	0.394	gaA	CCTACTATATGCCCAGAA	1.59E-11
18	P(T|AA)	0.251	aaT	CCTACTATATGCCCAGAAT	4.00E-12
19	P(T|AT)	0.323	atT	CCTACTATATGCCCAGAATT	1.29E-12

Scanning a sequence with a position-specific scoring matrix

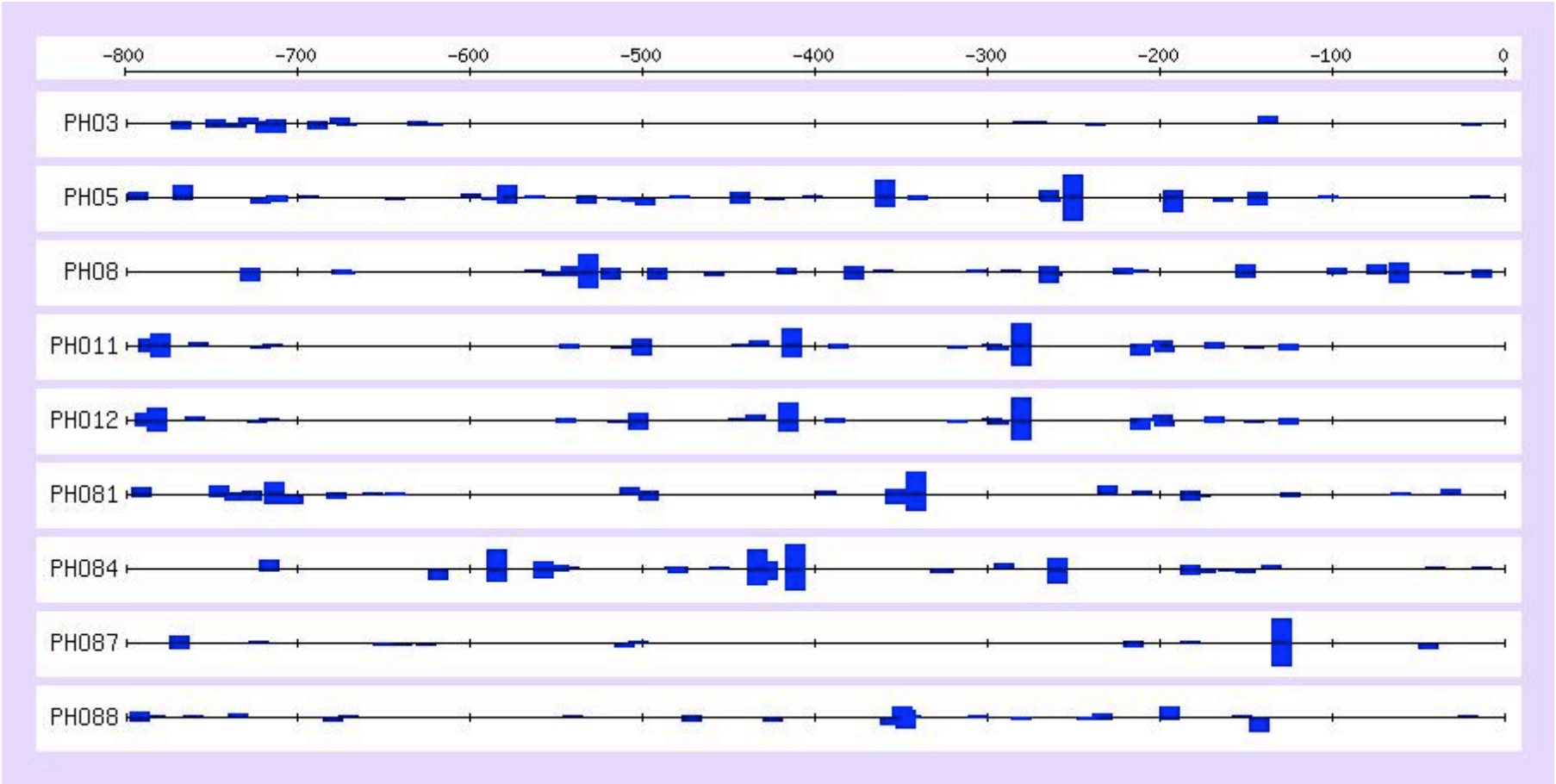
- **P(S|M)** probability for site S to be generated as an instance of the motif.
- **P(S|B)** probability for site S to be generated as an instance of the background.
- **W** weight, i.e. the log ratio of the two above probabilities.
  - A positive weight indicates that a site is more likely to be an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

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



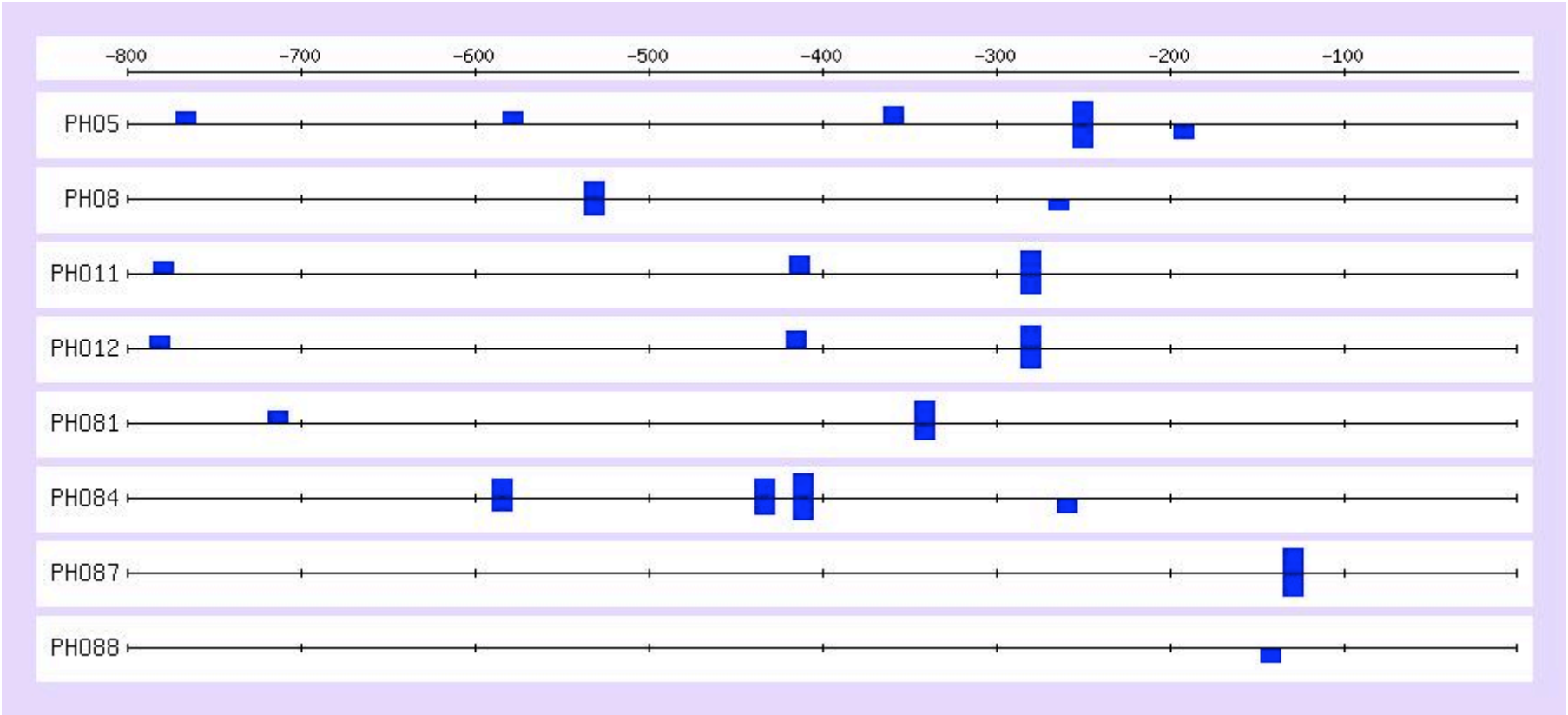
Matrix search : matching positions

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



Matrix search : threshold selection

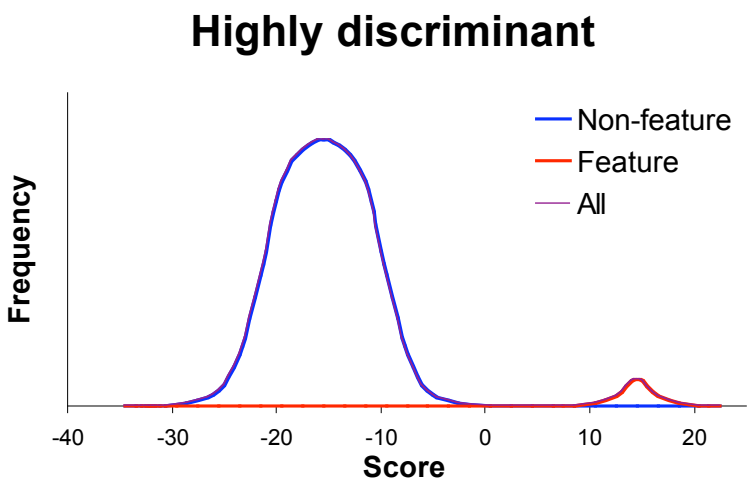
- Patser includes an option to automatically select a threshold on the basis of
  - the information content of the matrix
  - the length of the sequence to be scanned
- Another approach is to select the threshold on the basis of scores returned when the matrix is used to scan known binding sites for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

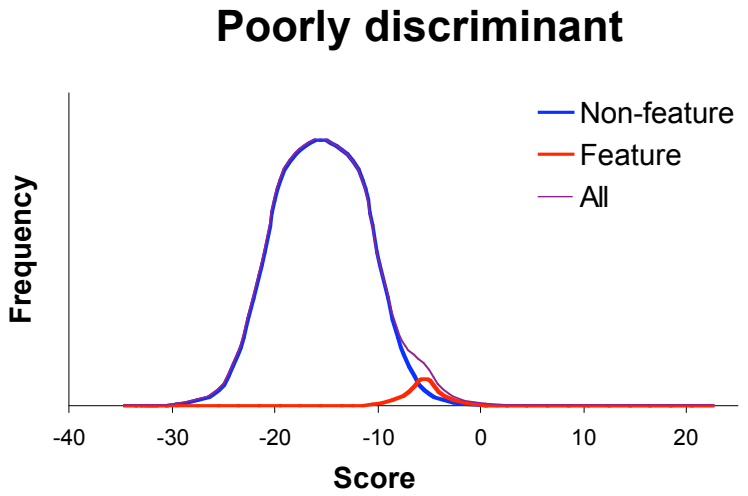


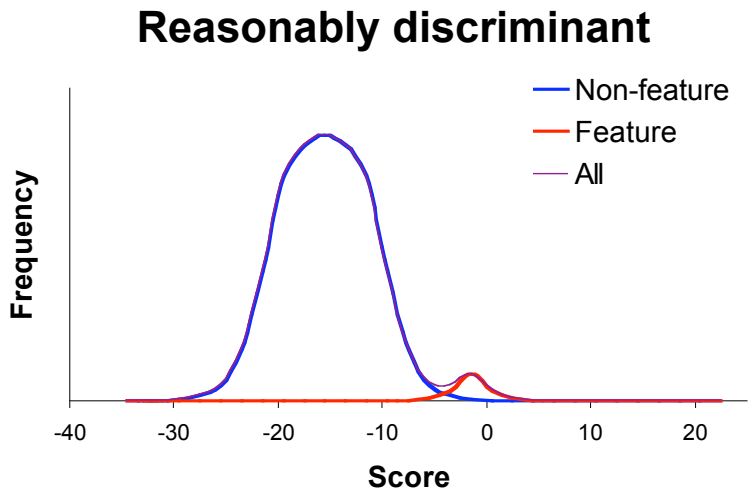
Matrix search

- 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

Discrimination power of a matrix







Regulatory sequence analysis

Theoretical distributions of weight scores

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

Score distribution: random expectation

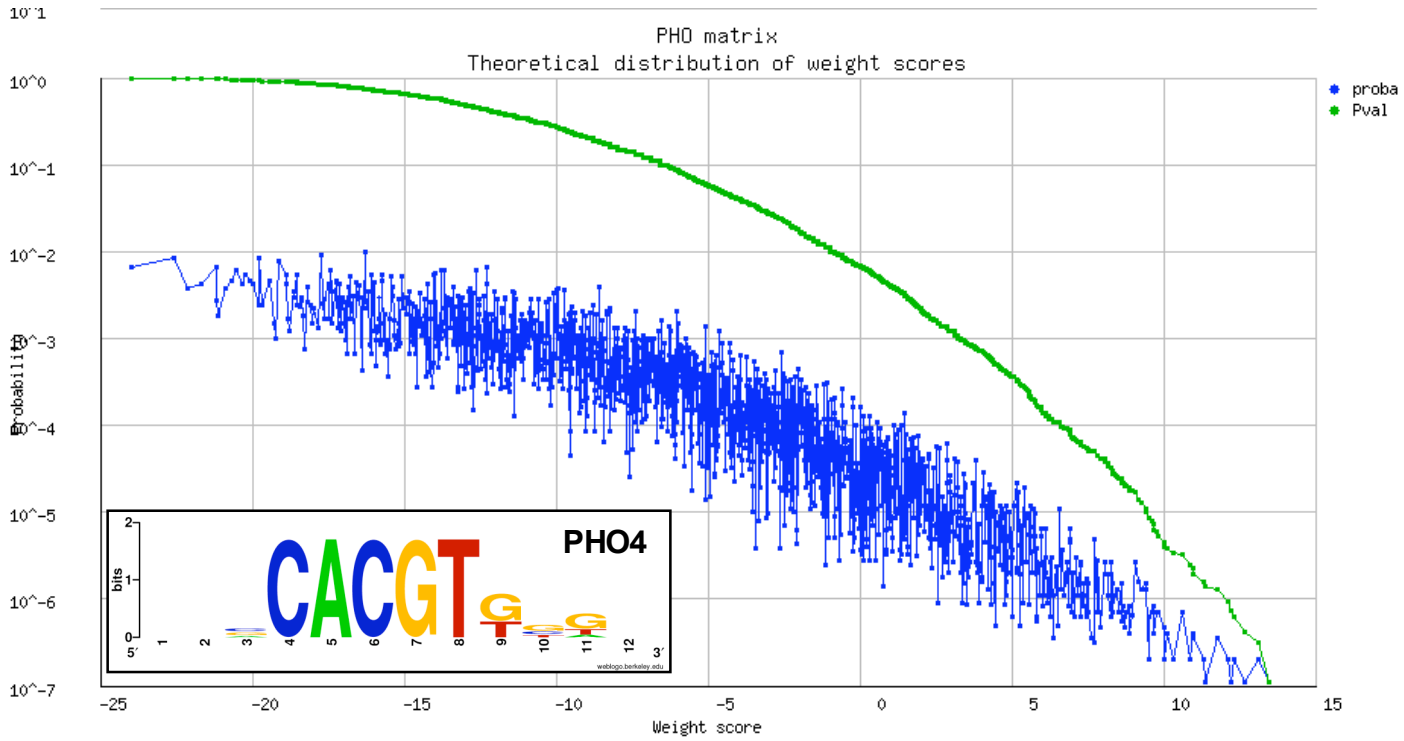
- The theoretical distribution of probabilities for position-weight matrices has been discussed in several articles.
  - Staden, R. (1989). Methods for calculating the probabilities of finding patterns in sequences. Comput Appl Biosci 5, 89-96.
  - Hertz, G. Z. & Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15, 563-77.
- The computation is based on the probability-generating function.
- This function can be used to compute the probability P(W) to obtain exactly a score value of W.
- Each position-weight matrix has its own probability-distribution.

$$G_j(x) = \sum f_i x w_{ij}$$

1. Bailey, T. L. & Gribskov, M. (1997). Score distributions for simultaneous matching to multiple motifs. *J Comput Biol* 4, 45-59.
2. Staden, R. (1989). Methods for calculating the probabilities of finding patterns in sequences. *Comput Appl Biosci* 5, 89-96.
3. Hertz, G. Z. & Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15, 563-77.

Theoretical distribution for the PHO matrix

- The figure below displays the theoretical distribution of scores for a PSSM obtained with consensus (J.Hertz) in upstream sequences of yeast PHO genes. Note that the Y axis is logarithmic.
- The theoretical distribution *P(S)* is quite erratic, because each possible value of score has its own probability, depending on
  - The actual weight values in the matrix
  - prior residue probabilities

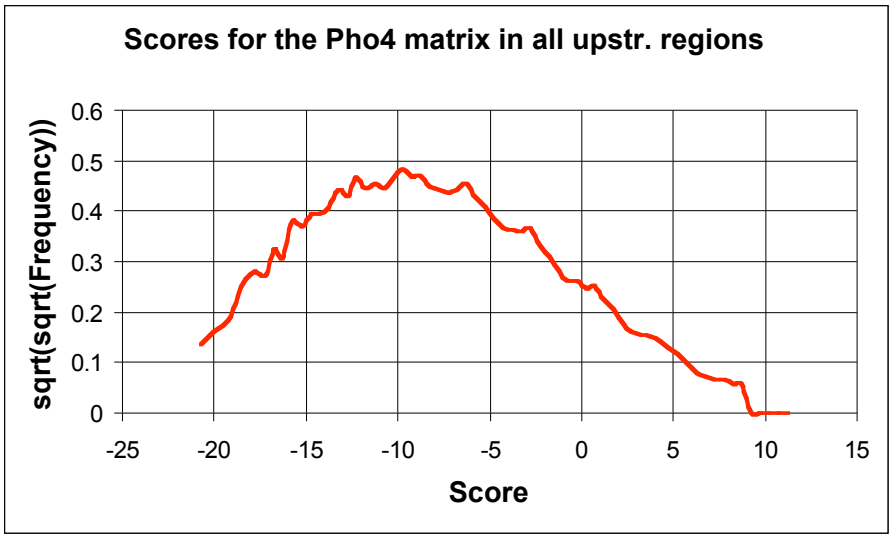


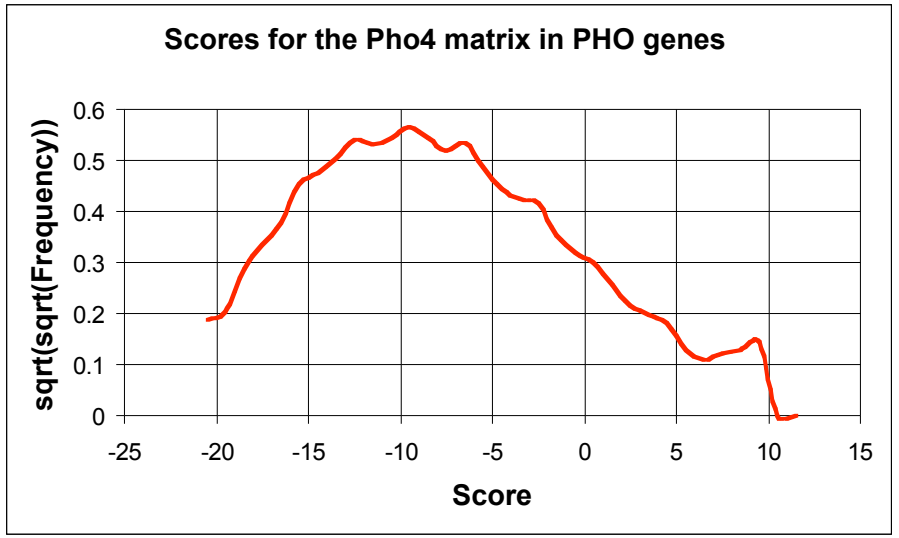
Regulatory sequence analysis

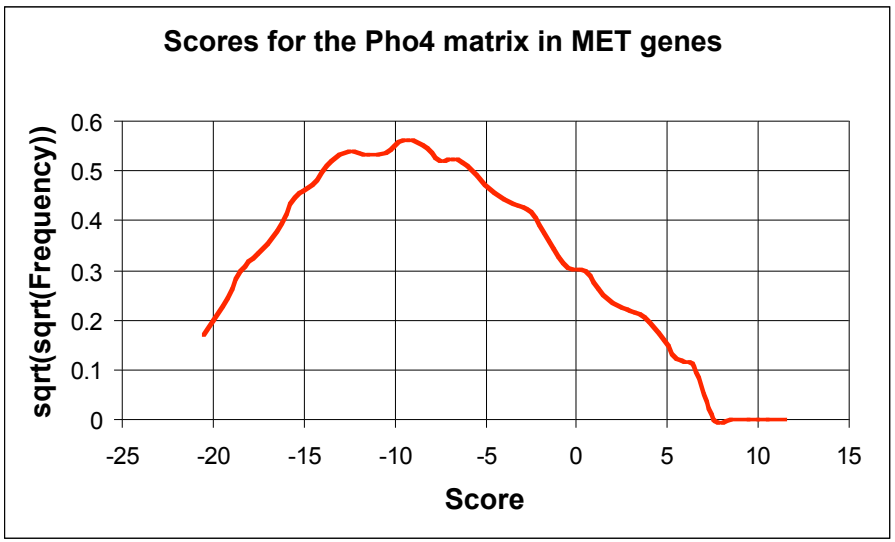
Genome-scale matrix-based pattern matching

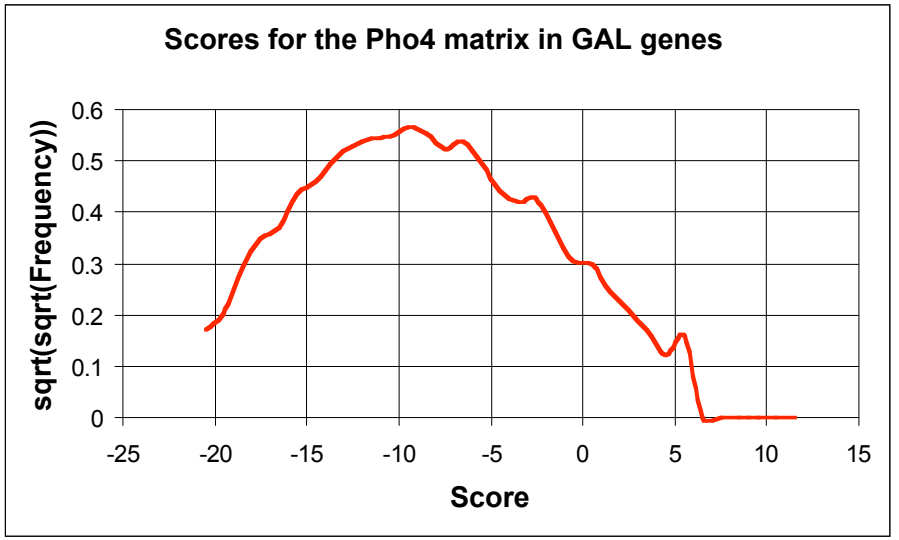
Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Pho4p matrix: score distribution in all upstream sequences









Genome-scale pattern matching: top scoring upstream sequences for the Pho4p matrix

ORF	strand	start	end	score	gene description
YMR255W	D	-326	-315	**10.2**	**hypothetical protein**
YKR050W	D	-744	-733	**9.83**	**TRK2; moderate-affinity potassium transport protein**
YKR048C	D	-442	-431	**9.83**	**NAP1; nucleosome assembly protein I**
YKR047W	D	-689	-678	**9.83**	**questionable ORF**
YKR031C	D	-597	-586	**9.83**	**SPO14; phospholipase D**
YCR037C	D	-135	-124	**9.83**	**PHO87; member of the phosphate permease family**
YAL038W	D	-525	-514	**9.83**	**CDC19; pyruvate kinase**
YML123C	D	-417	-406	**9.56**	**PHO84; high-affinity inorganic phosphate/H+ symporter**
YJL067W	D	-442	-431	**9.56**	**questionable ORF**
YCR051W	D	-298	-287	**9.56**	**weak similarity to ankyrins**
YBR296C	D	-326	-315	**9.56**	**strong similarity to phosphate-repressible phosphate permease**
YHR215W	D	-286	-275	**9.46**	**PHO12; secreted acid phosphatase**
YDR281C	D	-282	-271	**9.46**	**hypothetical protein**
YBR093C	D	-256	-245	**9.46**	**PHO5; repressible acid phosphatase precursor**
YAR071W	D	-286	-275	**9.46**	**PHO11; secreted acid phosphatase**
YNR050C	D	-516	-505	**9.35**	**LYS9; saccharopine dehydrogenase (NADP+, L-glutamate forming)**
YNL062C	D	-68	-57	**9.35**	**GCD10; translation initiation factor eIF3 RNA-binding subunit**
YJR059W	D	-246	-235	**9.19**	**PTK2; involved in polyamine uptake**
YHR107C	D	-506	-495	**9.19**	**CDC12; septin**
YGR233C	D	-347	-336	**9.19**	**PHO81; cyclin-dependent kinase inhibitor**
YFR040W	D	-81	-70	**9.19**	**SAP155; Sit4p-associated protein**
YJR067C	D	-505	-494	**9.08**	**YAE1; protein of unknown function**
YDR163W	D	-668	-657	**9.08**	**hypothetical protein**

Improving genome-scale predictions

- The prediction of single binding sites is notoriously noisy. However, one can try to combine different types of information in order to predict the regulation of a gene as a whole, rather than each potential binding site.
- Include as much biological information as possible
  - Site repetition (e.g. GATA boxes)
  - Combination of heterologous sites (e.g. Met4p and Met31p)
  - Information about site positions: distribution of position for known sites and impose a constraint on putative site positions
- Supervised classification: train a program with
  - a set of reliable positive sites (from known target genes)
  - a set of reliable negative sites
    (genes known not to respond to the factor)

Matching a sequence with a library of patterns

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

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$PACC_01          |      331 (-) |  1.000 |  0.897 | acgaGCCAagaaaagtt
  F$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$HSF_03           |      713 (-) |  1.000 |  0.859 | AGAAttgaaattttt
  F$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
```

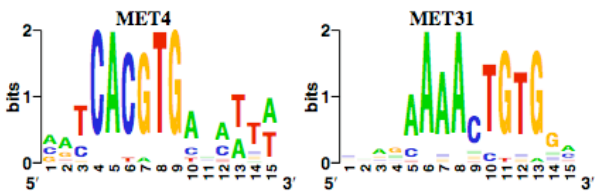
Regulatory sequence analysis

Cis-regulatory modules

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

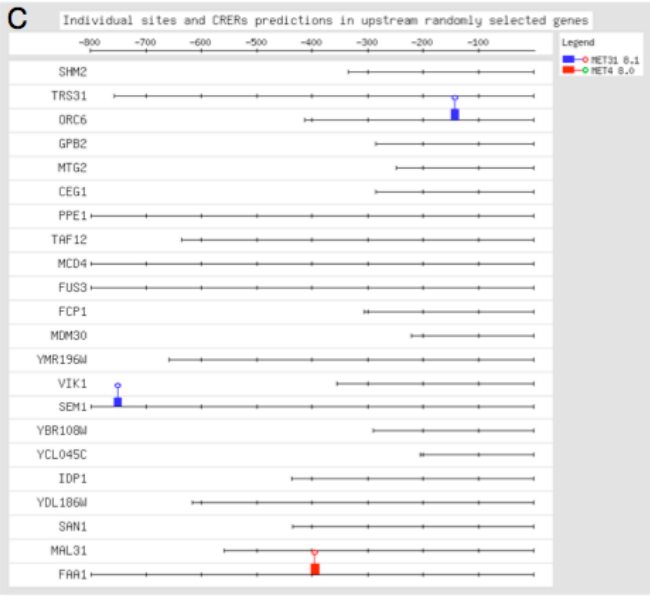
Cis-regulatory element enriched regions (CRERs) as putative cirs-regulatory modules (CRMs)

A



B

C



D

```
Background model
        Method          input
        Bernoulli model (order=0)
        Strand          insensitive
        Background pseudo-frequency     0.01
        Residue probabilities
                a       0.31711
                c       0.18289
                g       0.18289
                t       0.31711
Thresholds      lower   upper
        pval            NA      0.001
        score           0       NA
```

seq_id	ft_type	ft_name	strand	start	end	sequence	weight	proba_M	proba_B	Pval	ln_Pval	sig	rank	rank_pm
MET8	site	MET31	R	-429	-415	AGATAAAACTGCGGA	8.7	8.2e-06	1.2e-09	2.1e-05	-10.753	4.670	1	1
MET8	site	MET31	D	-183	-169	GAAAAAAAATGTGAA	8.6	4.0e-05	6.3e-09	2.4e-05	-10.649	4.625	2	2
MET8	site	MET4	R	-215	-201	TAACACGTGAAATTA	6.6	3.4e-06	3.6e-09	9.3e-05	-9.279	4.030	3	1
MET8	site	MET4	D	-212	-198	TTTCACGTGTTATAA	4.2	2.6e-07	3.6e-09	4.7e-04	-7.668	3.330	4	2
MET8	site	MET31	R	-260	-246	ATAAAACACTTTGAA	4.9	1.0e-06	6.3e-09	5.0e-04	-7.601	3.301	5	3
MET8	site	MET31	D	-80	-66	ATAAAAGGCTGTGCC	4.8	9.4e-08	7.0e-10	5.4e-04	-7.533	3.271	6	4

Regulatory sequence analysis

Old slides

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

Pho4p matrix : score distribution in upstream regions of PHO and MET genes