Matrices and Pattern matching

Morgane Thomas-Chollier

Computational systems biology - IBENS

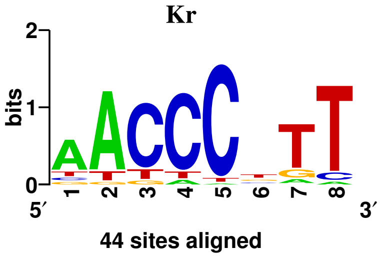
mthomas@biologie.ens.fr

M2 – Computational analysis of cis-regulatory sequences 2015/2016

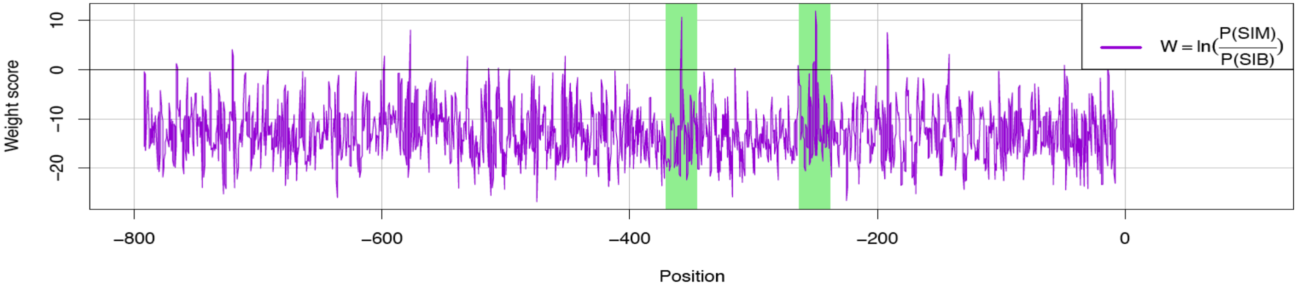
Denis Thieffry, Jacques van Helden and Carl Herrmann kindly shared some of their slides.

Aim of the course

1 - Understand what is a motif and its various representations



2 – Understand what is a pattern matching problem



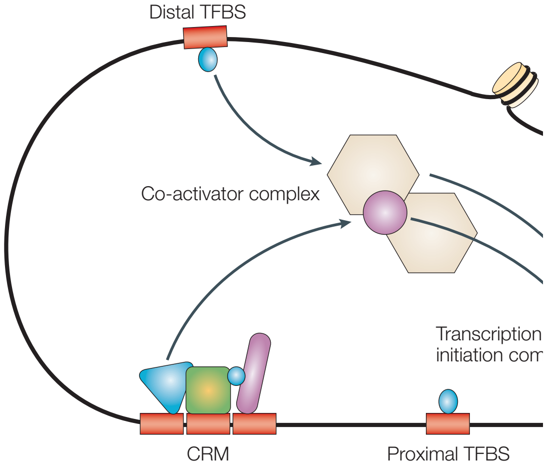
3 – Pattern matching approaches

1 - Understand what is a motif and its various representations

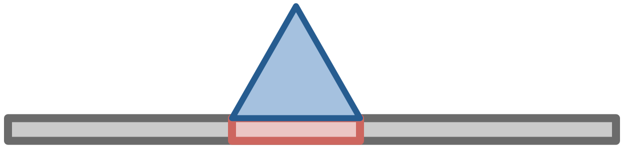
- From alignment to motif

- Motif descriptors

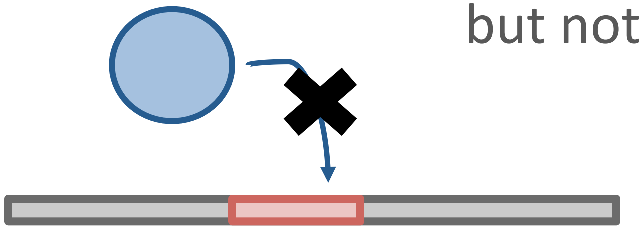
Transcription factor specificity



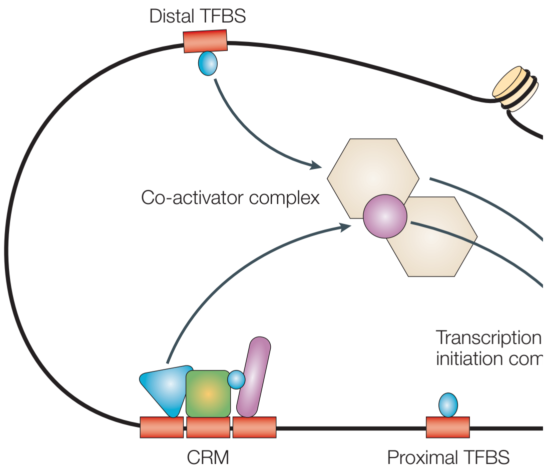
How do TF « know » where to bind DNA ?



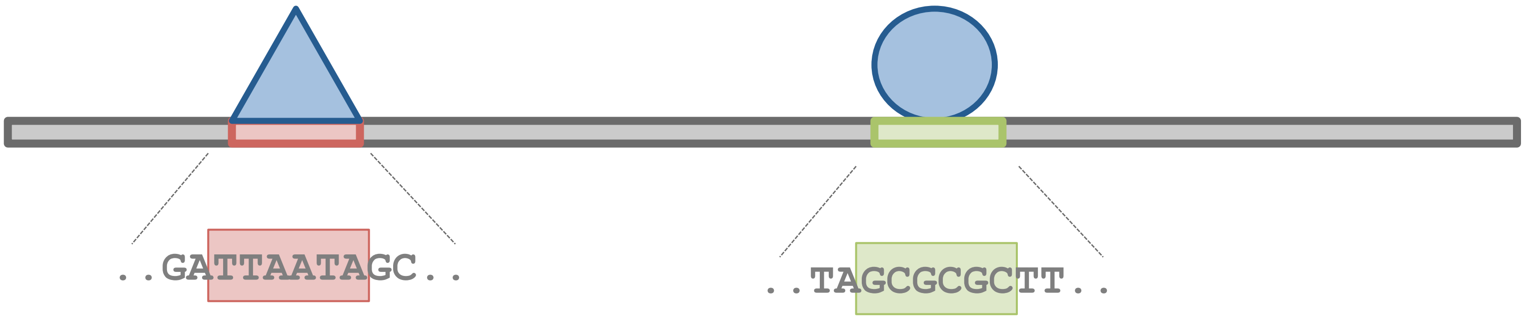
but not



Transcription factor specificity

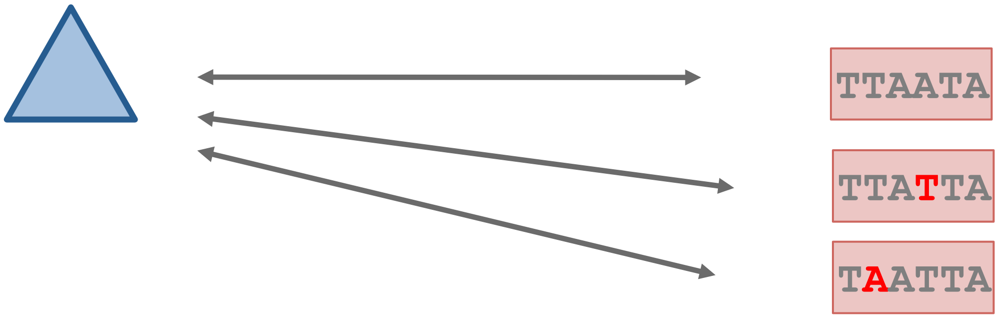


How do TF « know » where to bind DNA ?



TF recognize TFBS with specific DNA sequences

TFBS are degenerate

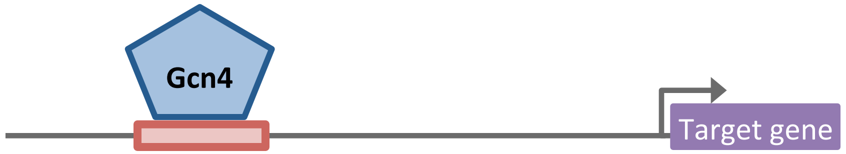


TFBSs are *degenerate*:
a given TF is able to bind DNA on TFBSs with different sequences

Problem : How can we model/describe the binding specificity of each TF ?
(for further usage by programs)

Binding specificity of a given TF

TF= Gcn4 (yeast transcriptional activator of genes involved in the biosynthesis of amino acids)



5'– TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAG**AAAAGAGTCA**GACATCGAAACATACAT ***…HIS7***

5'– ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCG**AAATGACTCA**ACG ***…ARO4***

5'– CACATCCAACGAATCACCTCACCGTTATCG**TGACTCACTT**TCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT ***…ILV6***

5'– TGCGAAC**AAAAGAGTCA**TTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC ***…THR4***

5'– ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATA**TGACTCATCC**CGAACATGAAA ***…ARO1***

5'– ATTGAT**TGACTCATTT**TCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAAATAAATAA ***…HOM2***

5'– GGCGCCACAGTCCGCGTTTGGTTATCCGGC**TGACTCATTCTGACTCTTTT**TTGGAAAGTGTGGCATGTGCTTCACACA ***…PRO3***

1	2	3	4	5	6	7	8	9	10
A	A	A	A	**G**	**A**	G	**T**	**C**	**A**
A	A	A	T	**G**	**A**	C	**T**	**C**	**A**
A	A	G	T	**G**	**A**	G	**T**	**C**	**A**
A	A	A	A	**G**	**A**	G	**T**	**C**	**A**
G	G	A	T	**G**	**A**	G	**T**	**C**	**A**
A	A	A	T	**G**	**A**	G	**T**	**C**	**A**
G	A	A	T	**G**	**A**	G	**T**	**C**	**A**
A	A	A	A	**G**	**A**	G	**T**	**C**	**A**

Multiple alignment

(! Orientation !)

DNA Binding motif

- *What is a motif ?*

- Sequence motifs are **short, recurring patterns** in DNA that are presumed to have **a biological function**. (D'haeseleer, Nature Biotechnology, 2006)

- **Common properties** shared by a group of biologically-related sequences (Bucher, Comput chem, 1996)

- In the context of transcriptional regulation:

For a given TF, the binding motif represents its binding specificity

A motif is an abstract concept !

- *How is a motif represented/described ?*

– Motif model or **motif descriptor** = representation of this motif

– a motif can be **represented synthetically** in various ways

1 - Understand what is a motif and its various representations

- From alignment to motif

- **Motif descriptors**

Motif descriptors

- **String-based**
  - Strict consensus
  - Degenerate consensus

- **Regular expressions**

- **Matrix-based**
  - Position-specific scoring matrices (PSSMs)

- **Sequence Logos**
- **Hidden Markov Models (HMM)**

String-based representation

- **Motif** is described as a **string** (=sequence)
  - **Consensus sequence** : derived from the collection of binding sites by taking the predominant letter at each position of the motif

TF= Gcn4

1	2	3	4	5	6	7	8	9	10
A	A	A	A	**G**	**A**	G	**T**	**C**	**A**
A	A	A	T	**G**	**A**	C	**T**	**C**	**A**
A	A	G	T	**G**	**A**	G	**T**	**C**	**A**
A	A	A	A	**G**	**A**	G	**T**	**C**	**A**
G	G	A	T	**G**	**A**	G	**T**	**C**	**A**
A	A	A	T	**G**	**A**	G	**T**	**C**	**A**
G	A	A	T	**G**	**A**	G	**T**	**C**	**A**
A	A	A	A	**G**	**A**	G	**T**	**C**	**A**

AAAt**GA**G**TCA** — **Strict consensus**: only ATGC alphabet

RAAw**GA**G**TCA** — **Degenerate consensus**: IUPAC code for ambiguous nucleotides

IUPAC ambiguous nucleotide code

A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
R	**A or G**	**puRine**
Y	C or T	pYrimidine
W	**A or T**	**Weak hydrogen bonding**
S	G or C	Strong hydrogen bonding
M	A or C	aMino group at common position
K	G or T	Keto group at common position
H	A, C or T	not G
B	G, C or T	not A
V	G, A, C	not T
D	G, A or T	not C
N	G, A, C or T	aNy

A **consensus** looks like a sequence but is not a TFBS (there is actually no TF recognizing this sequence, except in some cases of strict consensus) !!!

This is a **motif representation !!!**

Consensus sequence

😊 Simple and synthetic representation

TF= Gcn4

1	2	3	4	5	6	7	8	9	10
A	A	A	A	G	A	G	T	C	A
A	A	A	T	G	A	C	T	C	A
A	A	G	T	G	A	G	T	C	A
A	A	A	A	G	A	G	T	C	A
G	G	A	T	G	A	G	T	C	A
A	A	A	T	G	A	G	T	C	A
G	A	A	T	G	A	G	T	C	A
A	A	A	A	G	A	G	T	C	A

AAATGAGTCA

RAAwGAGTCA

☹ strict consensus: loss of information about non-predominant letters

☹ degenerate consensus: loss of information about the most frequent letter

Hands on !

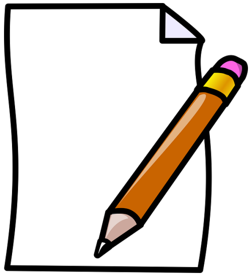
TF= Meis
(from various vertebrates)

TGACAA
TGACAG
TGATGG
TGACAA
TGGCAG
TGATTG
TGACAG
TGACAG

IUPAC ambiguous nucleotide code

A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
R	A or G	puRine
Y	C or T	pYrimidine
W	A or T	Weak hydrogen bonding
S	G or C	Strong hydrogen bonding
M	A or C	aMino group at common position
K	G or T	Keto group at common position
H	A, C or T	not G
B	G, C or T	not A
V	G, A, C	not T
D	G, A or T	not C
N	G, A, C or T	aNy

- **How many positions in this motif ?**
- **Write the strict consensus and degenerate consensus with IUPAC code**



Motif descriptors

- **String-based**
  - Strict consensus
  - Degenerate consensus

- **[Regular expressions]**

- **Matrix-based**
  - Position-specific scoring matrices (PSSMs)

- **Sequence Logos**
- **Hidden Markov Models (HMM)**

Matrix-based representation

TF= Gcn4

```
1 2 3 4 5 6 7 8 9 10
A A A A G A G T C A
A A A T G A C T C A
A A G T G A G T C A
A A A A G A G T C A
G G A T G A G T C A
A A A T G A G T C A
G A A T G A G T C A
A A A A G A G T C A
A A A T G A G T C A
R A A w G A G T C A
```

Count matrix

A	6	7	7	3	0	8	0	0	0	8
C	0	0	0	0	0	0	1	0	8	0
G	2	1	1	0	8	0	7	0	0	0
T	0	0	0	5	0	0	0	8	0	0

count matrix : indicates the number of times each nucleotide is found at each position of the motif.

😊 **More expressive than consensus sequence**
Keeps information on all nucleotides

Hands on !

TF= Meis

(from various vertebrates)

```
TGACAA
TGACAG
TGATGG
TGACAA
TGGCAG
TGATTG
TGACAG
TGACAG
```

- **Construct the count matrix for this TF**

Matrix-based representation

Count matrix (Krüppel matrix)

Residue i \position j	1	2	3	4	5	6	7	8
A	**30**	**37**	0	3	1	5	4	2
C	4	0	**35**	**37**	**41**	9	1	4
G	4	2	3	0	0	11	7	0
T	6	5	6	4	2	19	**32**	**38**
Sum	44	44	44	44	44	44	44	44

$$n_{i,j}$$

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

Frequency matrix (Krüppel matrix)

Residue\position	1	2	3	4	5	6	7	8
A	**0,68**	**0,84**	0,00	0,07	0,02	0,11	0,09	0,05
C	0,09	0,00	**0,80**	**0,84**	**0,93**	0,20	0,02	0,09
G	0,09	0,05	0,07	0,00	0,00	0,25	0,16	0,00
T	0,14	0,11	0,14	0,09	0,05	0,43	**0,73**	**0,86**
Sum	1	1	1	1	1	1	1	1

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

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

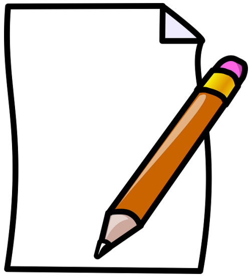
Reference: Hertz (1999). Bioinformatics 15:563-577.

Hands on !

TF= Meis
(from various vertebrates)

```
TGACAA
TGACAG
TGATGG
TGACAA
TGGCAG
TGATTG
TGACAG
TGACAG
```

- **Construct the frequency matrix for this TF**



Motif descriptors

- **String-based**
  - Strict consensus
  - Degenerate consensus

- **[Regular expressions]**

- **Matrix-based**
  - Position-specific scoring matrices (PSSMs)

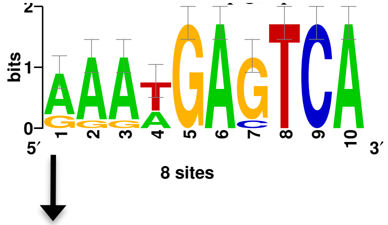
- **Sequence Logos**
- **Hidden Markov Models (HMM)**

Sequence Logo

TF= Gcn4

1	2	3	4	5	6	7	8	9	10
A	A	A	A	G	A	G	T	C	A
A	A	A	T	G	A	C	T	C	A
A	A	G	T	G	A	G	T	C	A
A	A	A	A	G	A	G	T	C	A
G	G	A	T	G	A	G	T	C	A
A	A	A	T	G	A	G	T	C	A
G	A	A	T	G	A	G	T	C	A
A	A	A	A	G	A	G	T	C	A
A	A	A	T	G	A	G	T	C	A
R	A	A	w	G	A	G	T	C	A

A	6	7	7	3	0	8	0	0	0	8
C	0	0	0	0	0	0	1	0	8	0
G	2	1	1	0	8	0	7	0	0	0
T	0	0	0	5	0	0	0	8	0	0



Logo

$$I_j = 2 - \left(-\sum_{i=1}^{A} f_{i,j} \log_2(f_{i,j})\right)$$

residue i , position j

- Graphical representation of a motif
- Each column represents one position of the motif
- The letters indicate which residues are found at a given position of the motif
- **Total height** of each column is proportional to the sequence conservation at this position (measured in bits)
- Maximum = 2 bits for a position that is perfectly conserved
- Indicates the **amount of information held by each position** of the motif
- The **height of each letter** is proportional to the **frequency of each residue** at a given position
- **Advantages:**
  - Easy identification of the most important positions of the motif

Schneider et al. Sequence logos: a new way to display consensus sequences. Nucleic Acids Research (1990) vol. 18 (20) pp. 6097-100

Motif descriptors

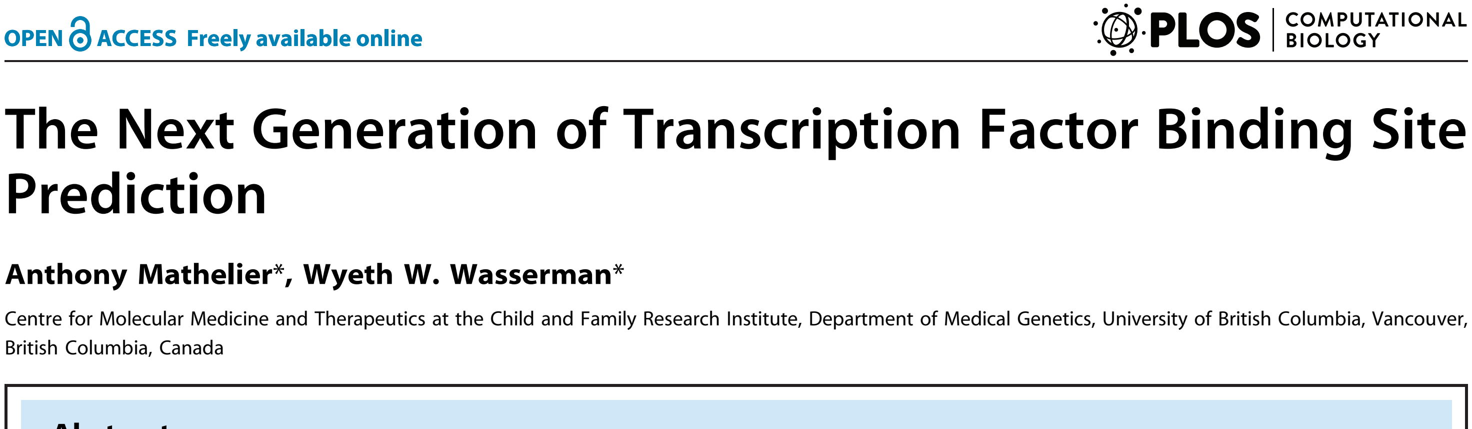
- **String-based**
  - Strict consensus
  - Degenerate consensus

- **[Regular expressions]**

- **Matrix-based**
  - Position-specific scoring matrices (PSSMs)

- **Sequence Logos**
- **Hidden Markov Models (HMM)**

HMM



OPEN ACCESS Freely available online

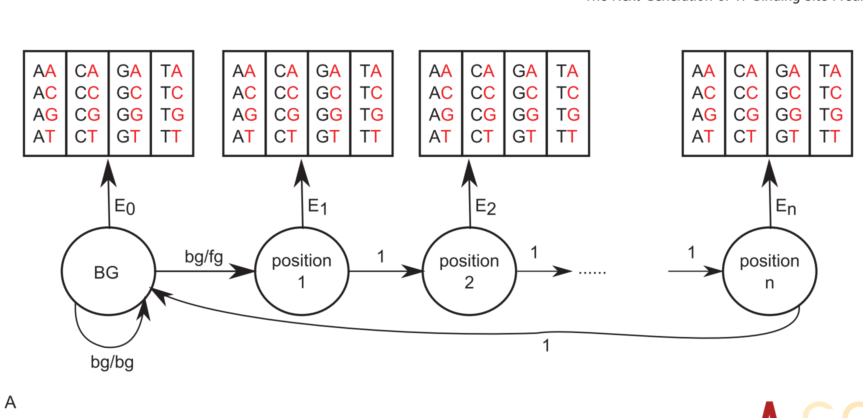
PLOS COMPUTATIONAL BIOLOGY

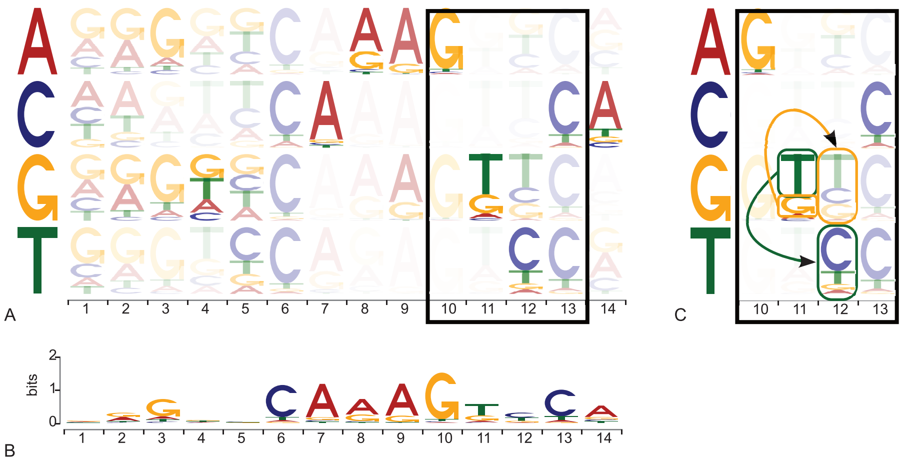
The Next Generation of Transcription Factor Binding Site Prediction

Anthony Mathelier*, Wyeth W. Wasserman*

Centre for Molecular Medicine and Therapeutics at the Child and Family Research Institute, Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada

Abstract



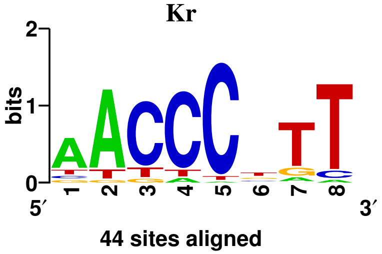


1 - Understand what is a motif and its various representations

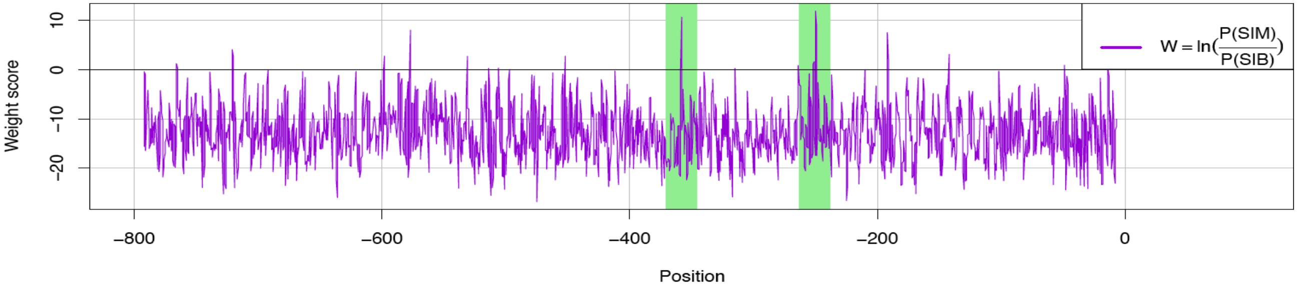
- The binding motif of a given transcription factor captures its DNA binding specificity.
- Among other ways to describe this motif, consensus sequences and count matrices are common and can be used by programs.
- The logo is a graphical descriptor to easily grasp the motif for a (human) eye.

Aim of the course

1 - Understand what is a motif and its various representations



2 – Understand what is a pattern matching problem



3 – Pattern matching approaches

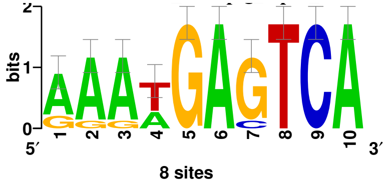
Pattern matching vs. Motif discovery (1)

Pattern matching

```
>genomic sequence
Tgcagagggatgatggacctttcgtctgtaggctcgtagaatggcacgcagtgggg
Aagagtgtgaccactcaatgctttgcagccgaaattgtggagaaaagtacggatag
Ttcccgtagttcgaggaagacataccatttattctccctccttgcgtaactaaatg
Gagaacattctaaagtgtcaccatcatatagacgatagataaccgatcgctgttcg
acttactcgggaaggacttggcacctttactccagtgagaacaacgtcccttagat
```

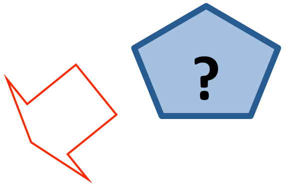
Is there a putative TFBS inside ?

At which position in the sequence ?



Problem : How can we search these motifs in sequences ?

Motif discovery

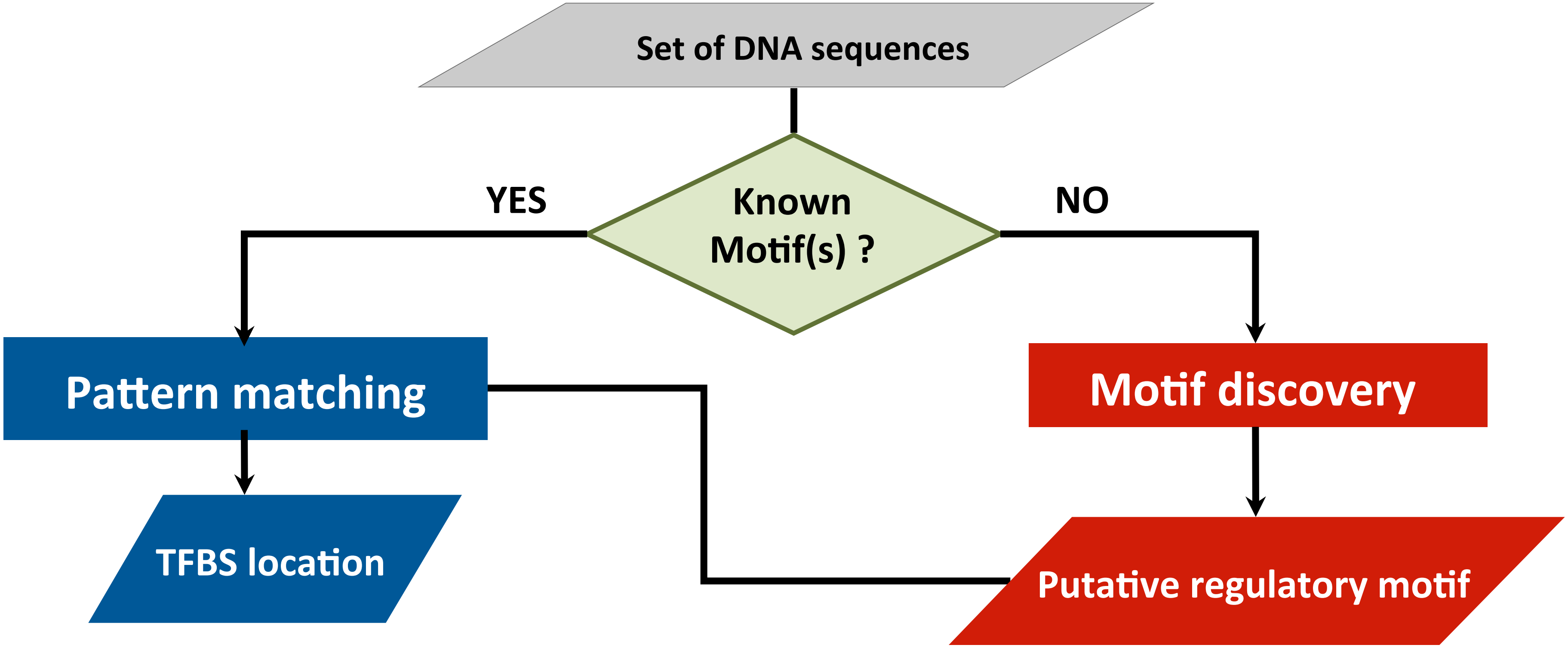


```
5’- TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT …HIS7
5’- ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG …ARO4
5’- CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT …ILV6
5’- TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC …THR4
5’- ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATATGACTCATCCCGAACATGAAA …ARO1
5’- ATTGATTGACTCATTTTCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA …HOM2
5’- GGCGCCACAGTCCGCGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTTGGAAAGTGTGGCATGTGCTTCACACA …PRO3
```

Co-expressed genes

Problem : If there is a common regulating factor, can we discover its motif only using these sequences ?

Pattern matching vs. Motif discovery (2)



Regulatory regions : characteristics

organism	bacteria	fungi	metazoan
location	upstream overlap. Initiation	upstream	upstream downstream within introns
distance range	-400 to +50 bp	-800 to -1 bp	from several kbs to several Mb !
position effect	often essential	often irrelevant	often irrelevant
strand	sensitive or symmetric	insensitive	insensitive
most common core	spaced pair of 3nt	~5-8 conserved bp	~5-8 conserved bp
repeated sites	rare	occasional	frequent
cis-regulatory modules (CRMs)			frequent

Computational detection of TFBS: Challenges

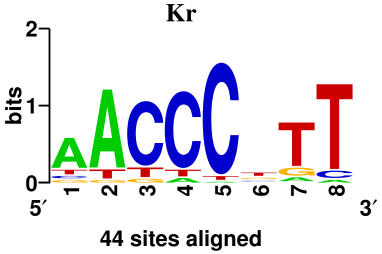
- Very **short** sequences (5 – 20 bp)
- TFBS are searched in sometimes **very large genomic sequences** (few Mb to Gb for vertebrate sequences !)
- TFBS are sometimes **far** their target genes
- TFBSs are **degenerate**: a given TF is able to bind DNA on TFBSs with different sequences

⚠ A computational **prediction** of a TFBS does **not imply that it is functional** biologically.

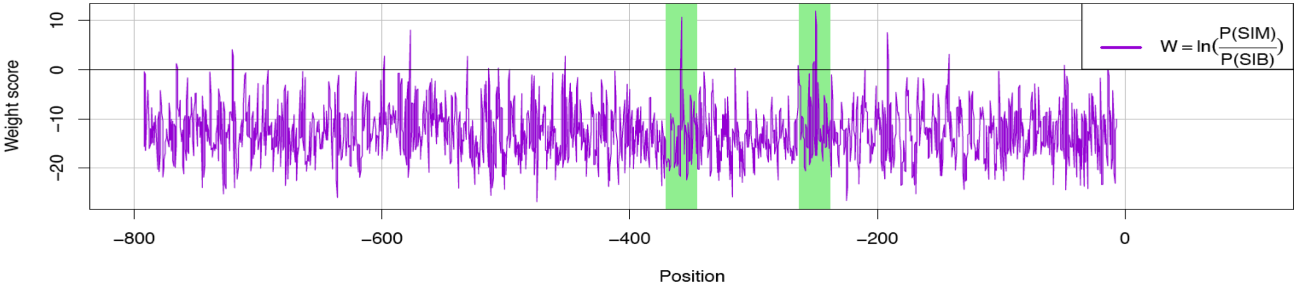
- Sequences identical to a TFBS are also found **at random sites** in genomes !
- On the contrary, a predicted TFBS shown not to be functional may be an actual TFBS in other **biological conditions**…

Aim of the course

1 - Understand what is a motif and its various representations



2 – Understand what is a pattern matching problem

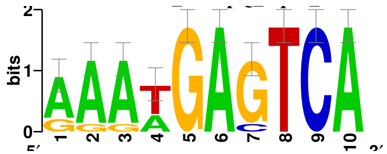


3 – Pattern matching approaches

Pattern matching

- String-based => motif = consensus
- Matrix-based => motif = count matrix
- Statistical evaluation of the results

Pattern-matching: principle



Gcn4 motif

INPUT

```
>sequence
Tgcagaggga
Aagagtgtga
Ttcccgtagt
Gagaacattc
acttactcggg
```

sequence of interest

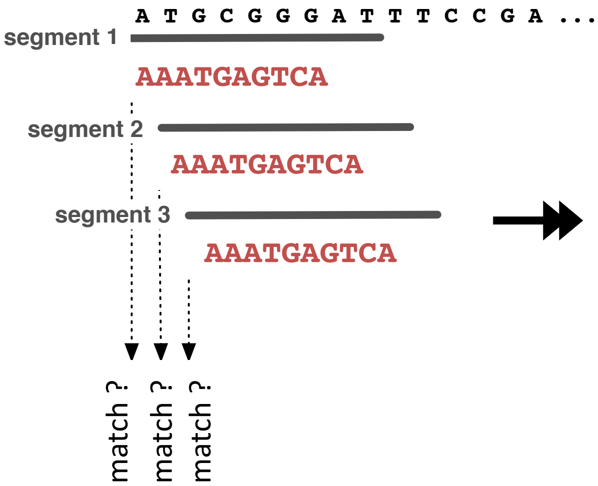
AAATGAGTCA

motif => consensus

Alignment of the motif with a sequence segment (segment length = motif length)

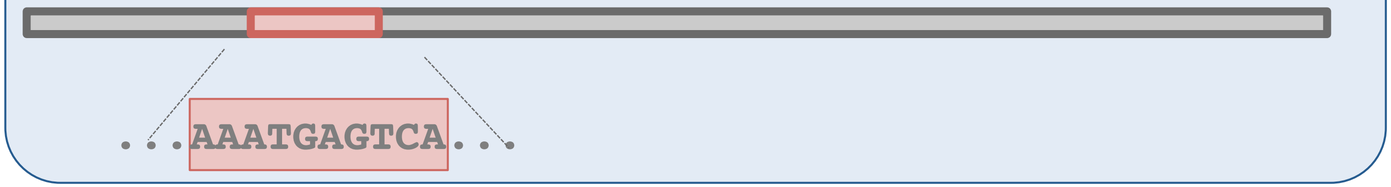
Is this position a match ?
=> Yes / No

Iterate to the next position in the sequence **« scanning »**



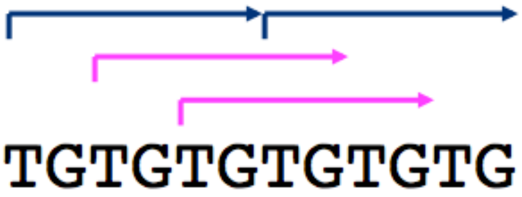
OUTPUT

positions of the matches on the sequence of interest



Problems to take into account

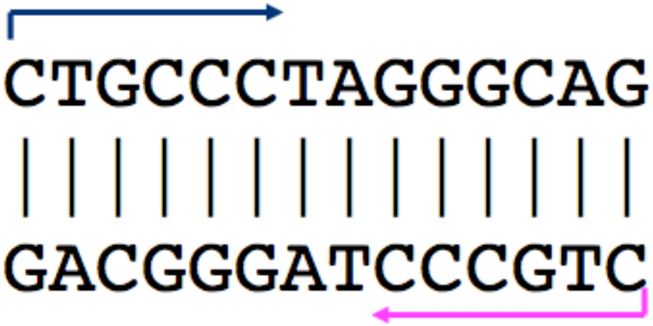
- treatment of self-overlap



TGTGTGTGTGTG

2 or 4 occurrences of TGTGTG ?

- Search on both strands ?



CTGCCCTAGGGCAG

GACGGGATCCCGTC

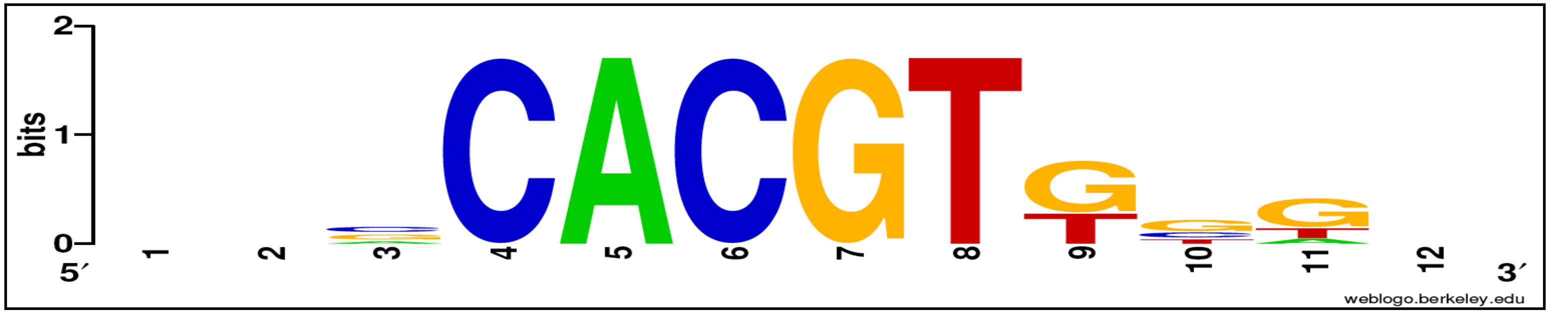
1 or 2 occurrences of CTGCCC ?

Pattern matching

- String-based => motif = consensus
- Matrix-based => motif = count matrix
- Statistical evaluation of the results

Regulatory motif described as a PSSM

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 motif of the yeast TF Pho4p (TRANSFAC matrix F$PHO4_01)

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*
$f_{i,j}$	*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	1+0.3	3	2	0	8	0	0	0	0	0	1	2
C	2+0.2	2	3	8	0	8	0	0	0	2	0	2
G	1+0.2	2	3	0	0	0	8	0	5	4	5	2
T	4+0.3	1	0	0	0	0	0	8	3	2	2	2
sum	8	8	8	8	8	8	8	8	8	8	8	8
Sum + pseudo	9											

- **Aim:**
  - Correct the **small-sample effect**
  - **Avoid "0" values** that will be problematic for computations with the PSSM
- **Principle**
  - Add a pseudo-count => "**fake**" additional site
  - If pseudo-count = 1, sum of occurences is thus +1
  - This value of "1" is **distributed among the four bases**
    - Equally: +0.25 to each letter
    - Prior : a residue-specific value is added

    Eg: yeast, upstream sequences: A =0.3 C=0.2 G=0.2 T=0.3

Corrected frequency matrix

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

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

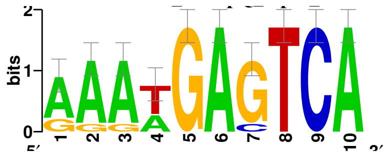
2nd option: pseudo-weight distributed according to residue priors

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

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.

Pattern-matching: principle



Gcn4 motif

INPUT

```
>sequence
Tgcagaggga
Aagagtgtga
Ttcccgtagt
Gagaacattc
acttactcggg
```

sequence of interest

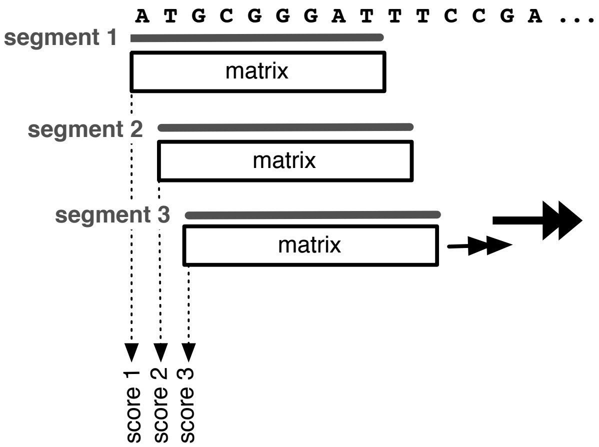
A	6773080008
C	0000001080
G	2110807000
T	0005000800

motif => matrix

Alignment of the motif with a sequence segment (segment length = motif length)

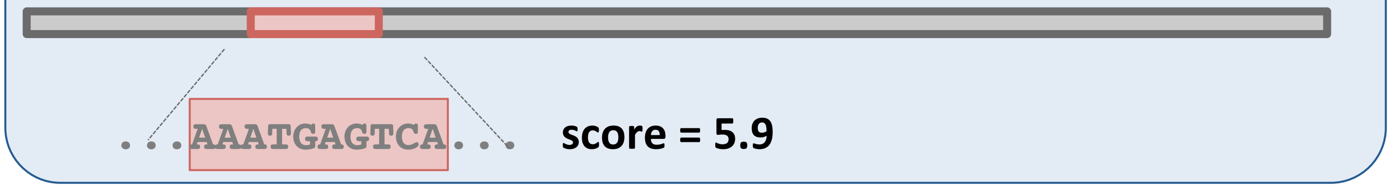
Is this position a match ? score calculation : Is the score above the threshold ?

Iterate to the next position in the sequence **« scanning »**



OUTPUT

positions of the matches on the sequence of interest **+ score**



score = 5.9

Probability of a sequence segment under the **matrix** model P(S|M)

>my_sequence_to_scan
ATGCGTAAAGCTAAAATTCTGTAAGACTAGAATCCAGGAGGCCAACTGTGATTGAGTTCTGAAAAATTGAGAGCCCTACTCCCCTCTC
TCACTTGTGGGAGCCCACTCAGGTCTGAAGTGCTCCCAGAGAACATGCCAGAATTAC....

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 a sequence segment under the **background model** P(S|B)

Pos	Prior
A	0.325
C	0.175
G	0.175
T	0.325

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

Sequence S	A	T	G	C	G	T	A	A	A	G	C	T
P(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**.
- One way to view the background model is to consider the matrix as what models the binding sites and the **background as what models non-binding sites**
- 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.

Assigning a score to a DNA sequence segment: the weight score

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

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

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.

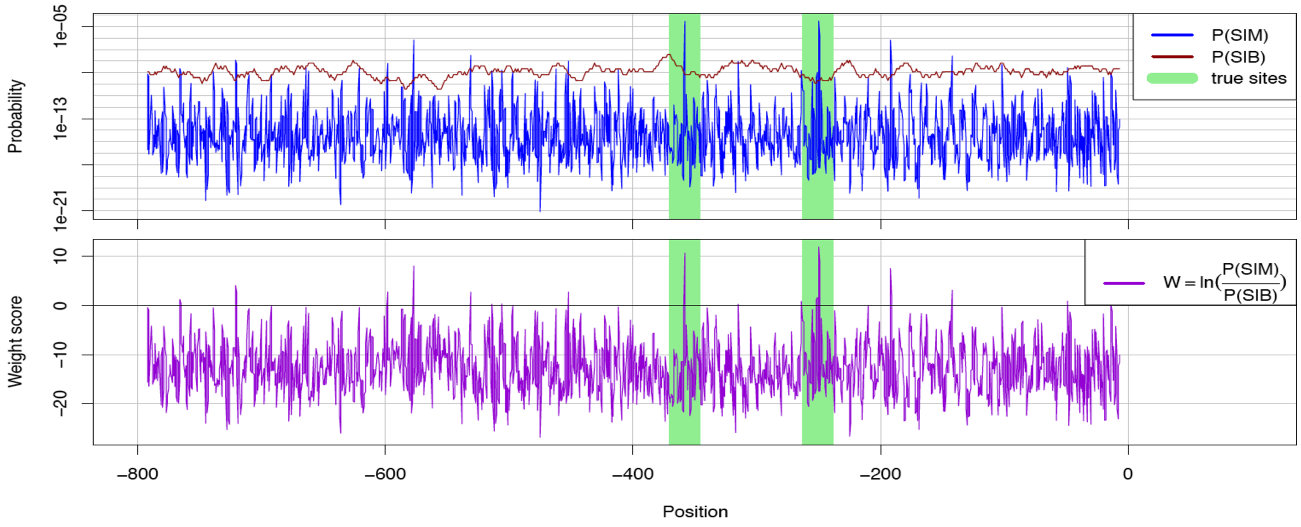
$$W_S = \ln\left(\frac{P(S \mid M)}{P(S \mid B)}\right)$$

P(S|M)= 5.32E-13

P(S|B) = 6.29E-8

W_S= -11,67

A positive weight indicates that a site **is more likely** to be an instance of the motif than of the background.



Sand, O., Turatsinze, J.V. and van Helden, J. (2008). Evaluating the prediction of cis-acting regulatory elements in genome sequences In Frishman, D. and Valencia, A. (eds.), Modern genome annotation: the BioSapiens network. Springer.

Hands on !

AATGAC

TF= Meis
(from various vertebrates)

a	0	0	7	0	6	2
c	0	0	0	6	0	0
g	0	8	1	0	1	6
t	8	0	0	2	1	0

- **Calculate the probability of the green sequence under the matrix model**

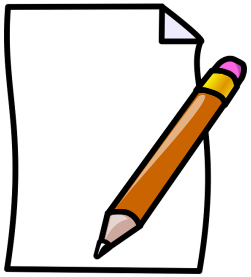
$$P(S \mid M) = \prod_{j=1}^{w} f'_{r_j j}$$

- **Calculate the probability of the green sequence under the background model pA=pT=0.2 pC=pG=0.3**

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

- **Calculate the weight score of this sequence**

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



Position-weight matrix (PWM)

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

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)]									

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

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

W_S weight of sequence segment S
$W_{i,j}$ weight of residue i at position j

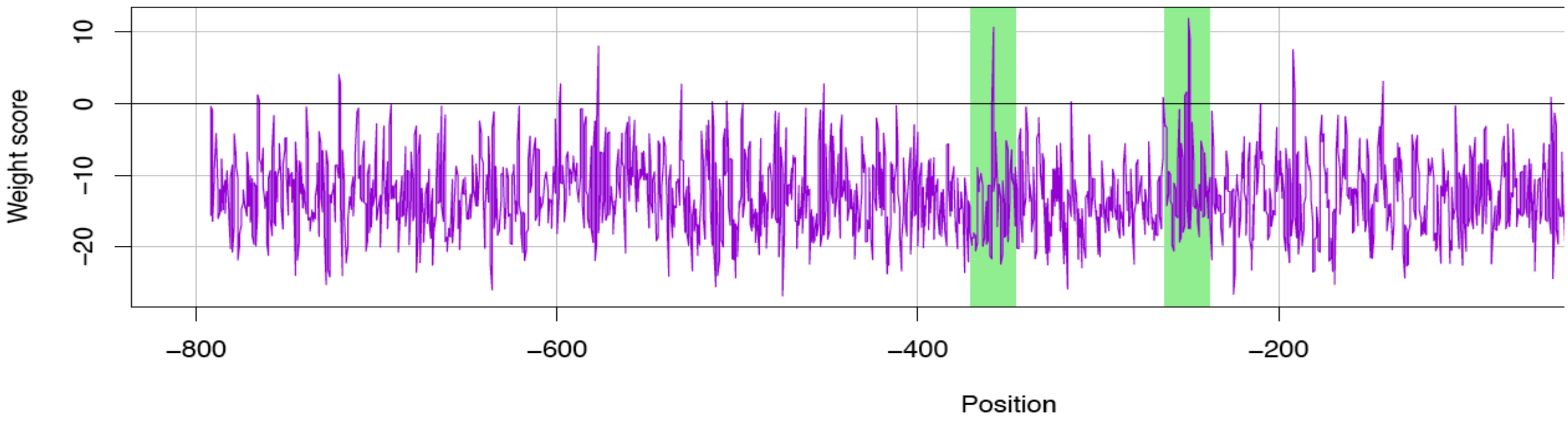
p_i prior residue probability for residue I
$f'_{i,j}$ corrected frequency of residue i at position j

$P(S|M)$ probability of the sequence segment, given the matrix
$P(S|B)$ probability of the sequence segment, given the background

Pattern matching

- String-based => motif = consensus
- Matrix-based => motif = count matrix
- Statistical evaluation of the results

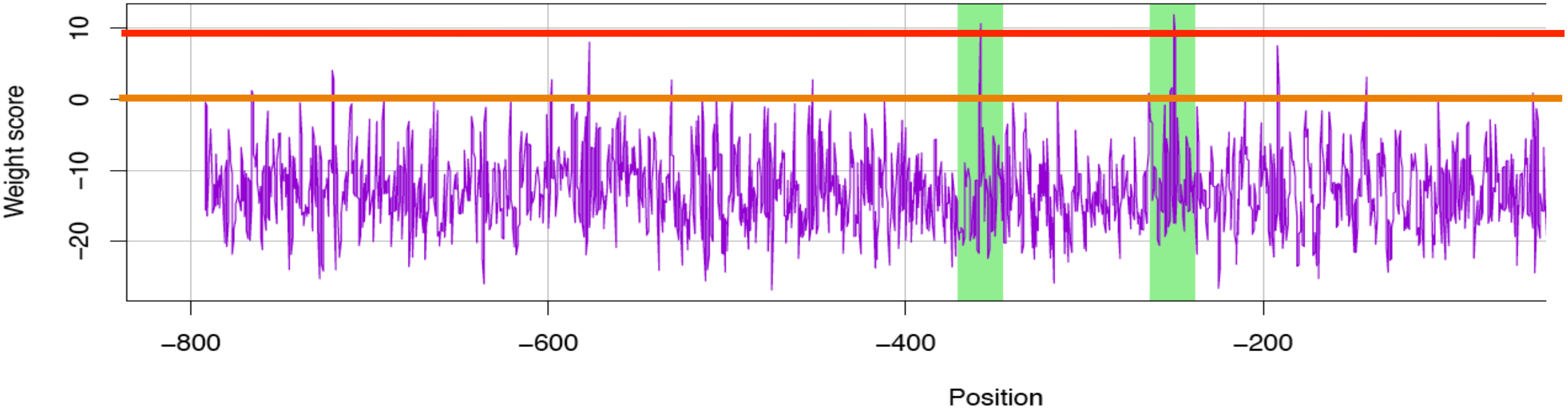
Pattern matching : a delicate compromise



- 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**.

Where to set the threshold ?

Pattern matching : a delicate compromise



- How to define the threshold ? There is a trade :
  - **stringent threshold**

    => high confidence in the predicted sites, but many real sites are missed

  - **loose threshold**

    => the real sites are drawn in a sea of false positive

Pattern matching : a delicate compromise

Predictions

Annotation

	Positive	Negative
Positive	True Positive	False negative
Negative	False Positive	True Negative

Total sites = TP+FN

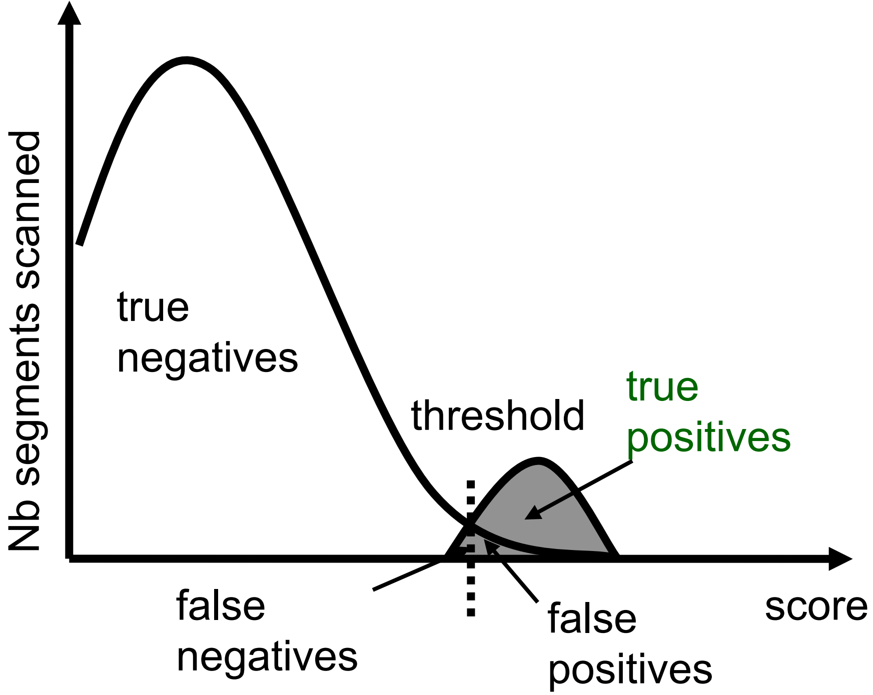
Total “hits”= TP+FP

Positive Predictive Value (PPV) **Selectivity** = Nb True Positives / Nb Total hits

Sensitivity = Nb True Positives / Nb Total sites

Trade between

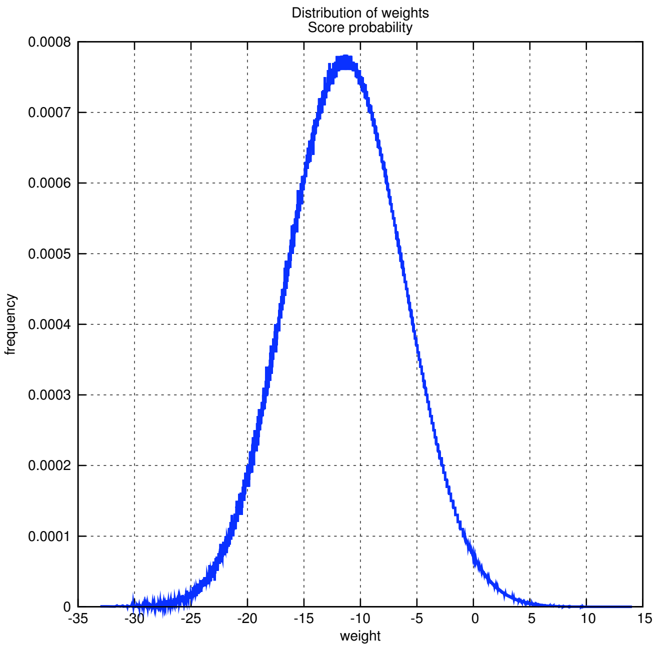
- **high selectivity ⇔ low sensitivity**
  => high confidence in the predicted sites, but many real sites are missed
- **low selectivity ⇔ high sensitivity**
  => the real sites are drawn in a sea of false positive



Score distributions and P-values

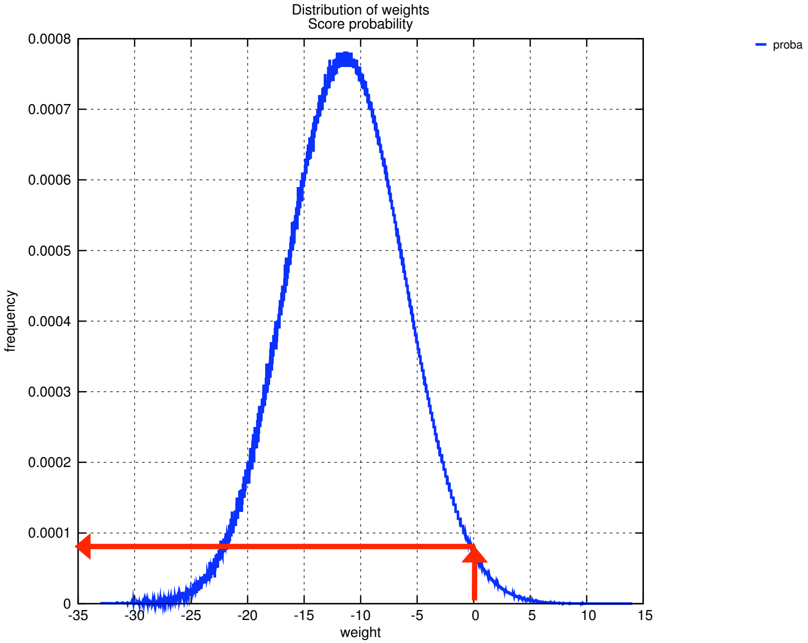
```
eve PSSM (length = 15)
a      |     0     5     3     6     1     2     7     0     0     8     4     2     3     1     0
c      |     8     0     4     2     3     3     0     0     1     1     0     5     0     6     7
g      |     1     4     2     1     0     3     2     0     0     0     4     2     6     2     0
t      |     0     0     0     0     5     1     0     9     8     0     1     0     0     0     2
```

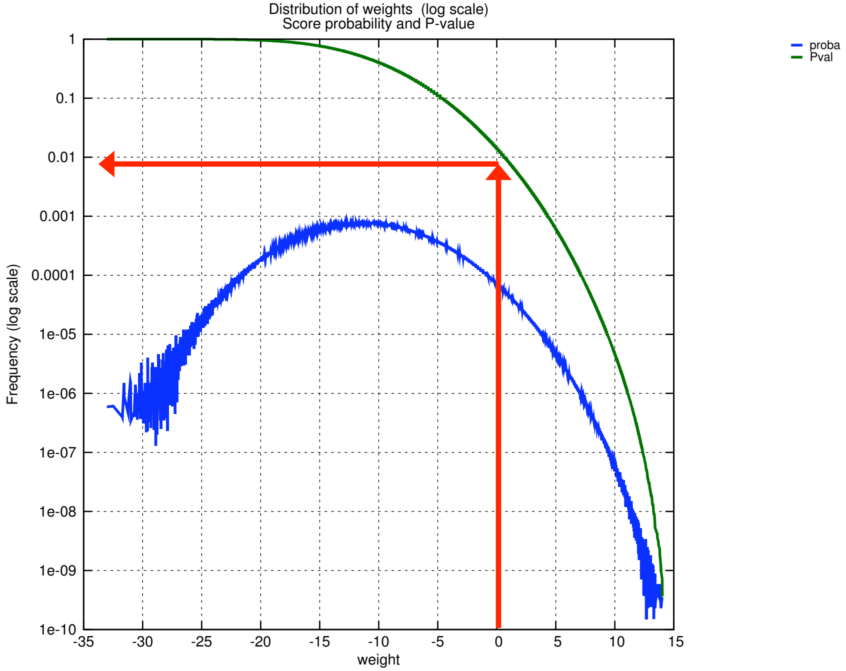
- **Imagine a virtual experience:**
  - Generate all possible words that can be scored with this PSSM : 4^{15} sequence segments of size 15 (ranging from `AAAAAAAAAAAAAAA` to `TTTTTTTTTTTTTTT`)
  - Given a background model, we score each segments
  - Obtain a list of 4^{15} scores => calculate the frequency of each score
  - Obtain the **theoretical distribution of scores**



- **In practice:**
  - This approach is not computionnally efficient and becomes impossible for large matrices
  - Directly calculate the distributions, without generating or scoring any sequence

Score distributions and P-values





Probability to observe **by chance** the **exact** score P(X=x)

eg: P(X=0) = 7.2 E-5

inverse cumulative distribution:

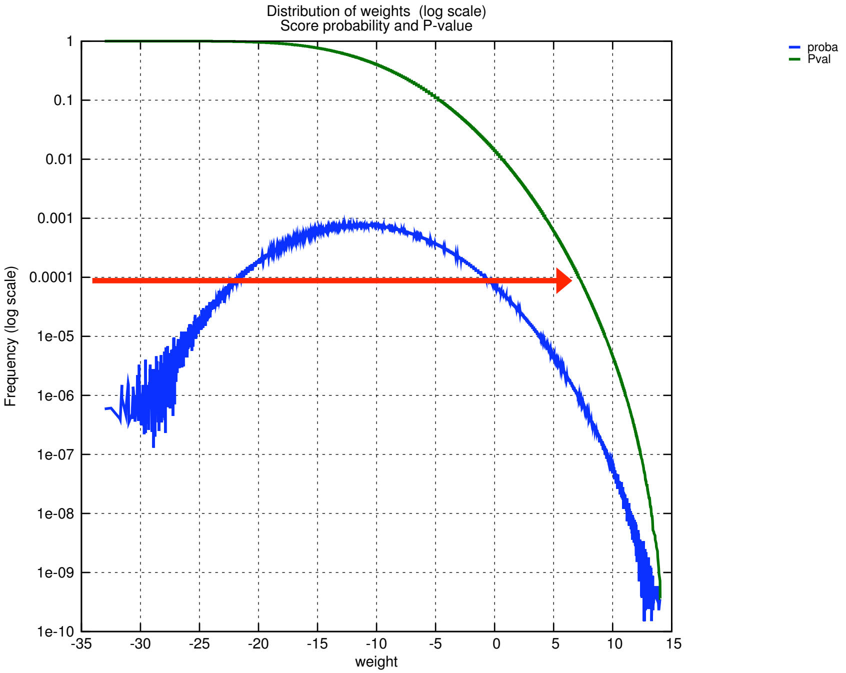
Probability to observe **by chance** a score of **a least x** P(X>=x)

eg: P(X>=0) = 1.4 E-2

- P(X>=x) is the **P-value** associated to score x
- The **P-value** is interpreted as **the risk of false predictions**

Setting a threshold on P-values

Threshold can be interpreted in terms of risk of false predictions

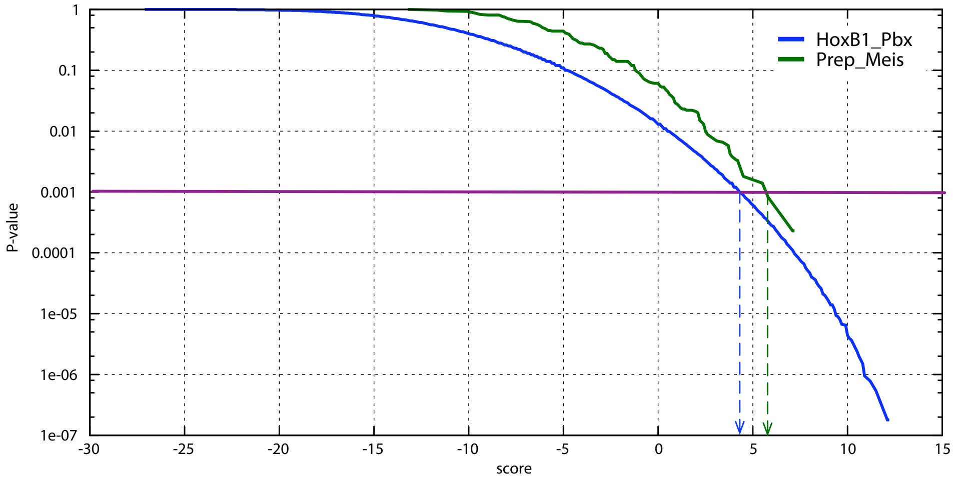


Set a threshold on P-value = **10 E-4**

=> Contrary to a threshold set on weight, this threshold can be interpreted:
we expect **1 false prediction** every **10 000 bp** if scanned on **one** strand
we expect **2 false predictions** every **10 000 bp** if scanned on **both** strands

Setting a threshold on P-values

Allows to work with multiple matrices



– A score in a given matrix **does not correspond to the same score** in another matrix

eg: depending on the size of the matrix, a score of 5 could be a very high or quite low score

– Using a **threshold on P-value** circumvents this issue

– **Tracing the theoretical distribution** helps to define an appropriate threshold

Strategies for pattern matching

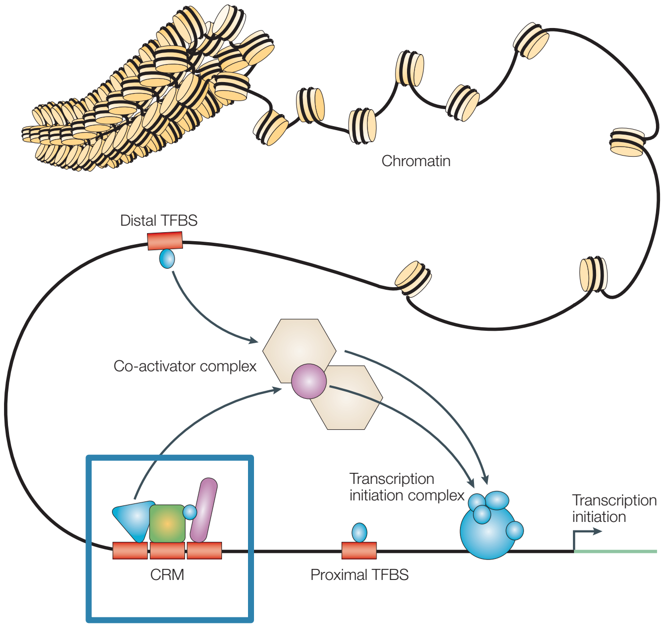
- Pattern-matching aims at finding putative TFBS (for which the binding motif is known) in DNA sequences
- If the motif is represented as a matrix, a weight score is calculated for all possible positions, but only the regions with a score higher than a threshold are considered as hits
- Results are highly dependent on this threshold. Choosing a threshold on a p-value allows to interpret it in terms of risk of false predictions.

Pattern-matching

Searching for regulatory modules

Cis-Regulatory Modules (CRMs)

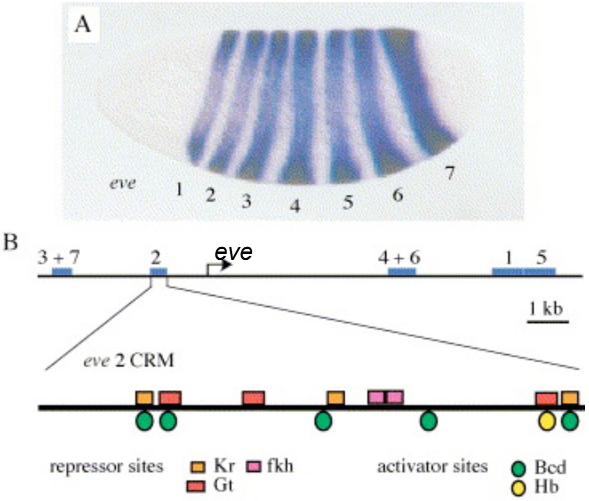
- What is a CRM ?
  - Various definitions and names (promoter modules, cis-regulatory clusters, composite elements)
  - Relatively small genomic region (hundreds of nucleotides) regrouping several binding sites, allowing a combined effect of multiple TFs on the expression of the target gene



Wasserman et al, Nat Rev Genet, 2004

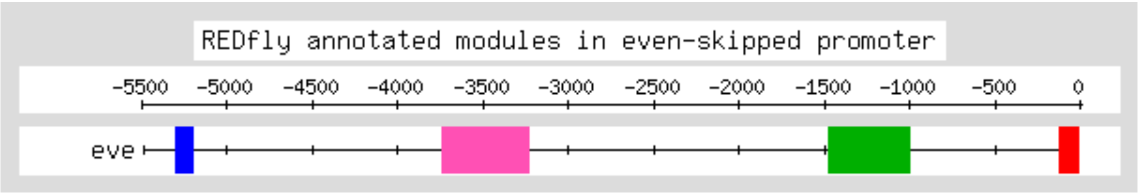
Cis-Regulatory Modules (CRMs)

- Example of CRM:
  - ***eve*** in *Drosophila melanogaster*
  - Multiple CRMs located upstream and downstream the *eve* gene, driving its expression in specific stripes in the embryo



Howard and Davidson, Dev Biol, 2004

- Some experimentally-detected CRMs are annotated in databases (ORegAnno, REDfly, Pazar,Transfac), along with their genomic coordinates (e.g. eve_stripe2 CRM is located on chromosome 2R from 5865266 to 5865750 in the Release 5 of the genome assembly)



mas
(autoregulatory element)

Stripe 3+7

Stripe 2

proximal
promoter

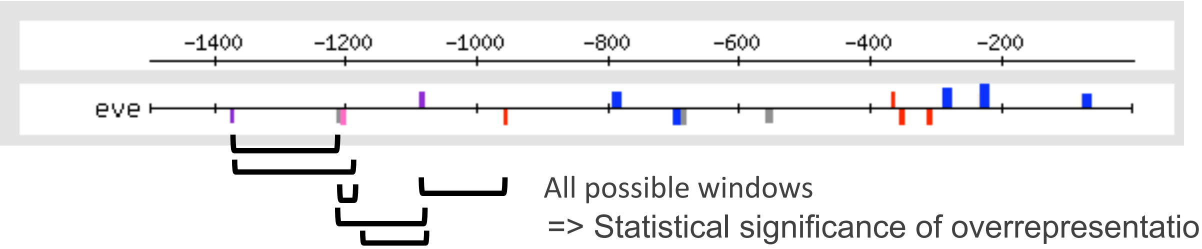
Detecting CRMs : principle

- Detection of regions containing **a higher density of predicted TFBS** than expected by chance => **Cis-Regulatory Enriched Region (CRER)**
- Various programs have been implemented to predict CRMs (Cluster-Buster, ModuleMiner).
- The RSAT program *matrix-scan* supports CRM prediction, by detecting regions enriched in cis-regulatory elements (CRERs).

- **Principle**

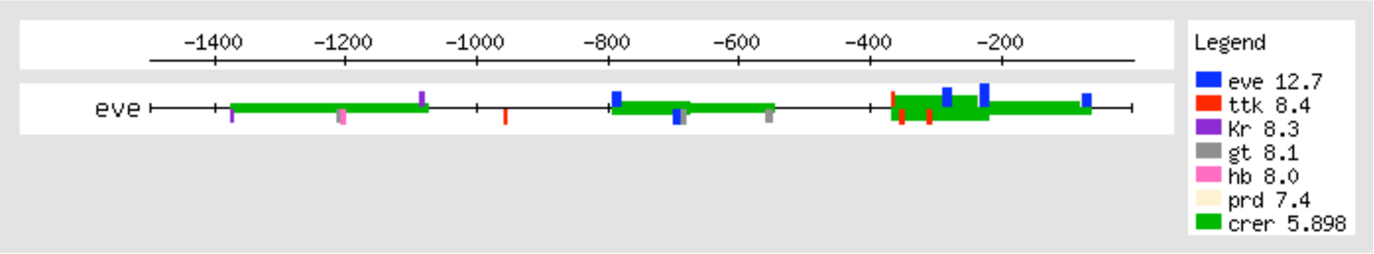
  The program (1) predicts all sites passing a threshold on P-value for each of the input matrices, and (2) detects regions (windows) having significantly more hits than expected by chance.

Site prediction
Threshold= 10^{-4}



All possible windows
=> Statistical significance of overrepresentation of predicted sites

CRER prediction



CRER prediction with matrix-scan

- Main features
  - Detection of homotypic (single motif) or heterotypic (distinct motifs) models.
  - No need to specify somewhat arbitrary constraints like the number of desired sites for each TF in a CRM, or the spacing between individual TFBS predictions.
  - All possible windows are tested within a user-speficied width range (e.g. from 30 to 300).
  - The enrichment is estimated by using the binomial statistics
  - The P-value estimates the risk of error when considering that a region contains more matches than expected by chance

$$P-value(y) = P(Y >= y) = \sum_{i=y}^{n} \binom{n}{i} P_\theta (1 - P_\theta)^{n-i}$$

y *nb of motifs occurences*
P_θ *user-selected threshold on individual site P-value*
n *nb of positions where a site can be predicted*

A threshold can be assigned on the significance of the CRER (only highly significant CRERs are thus returned)

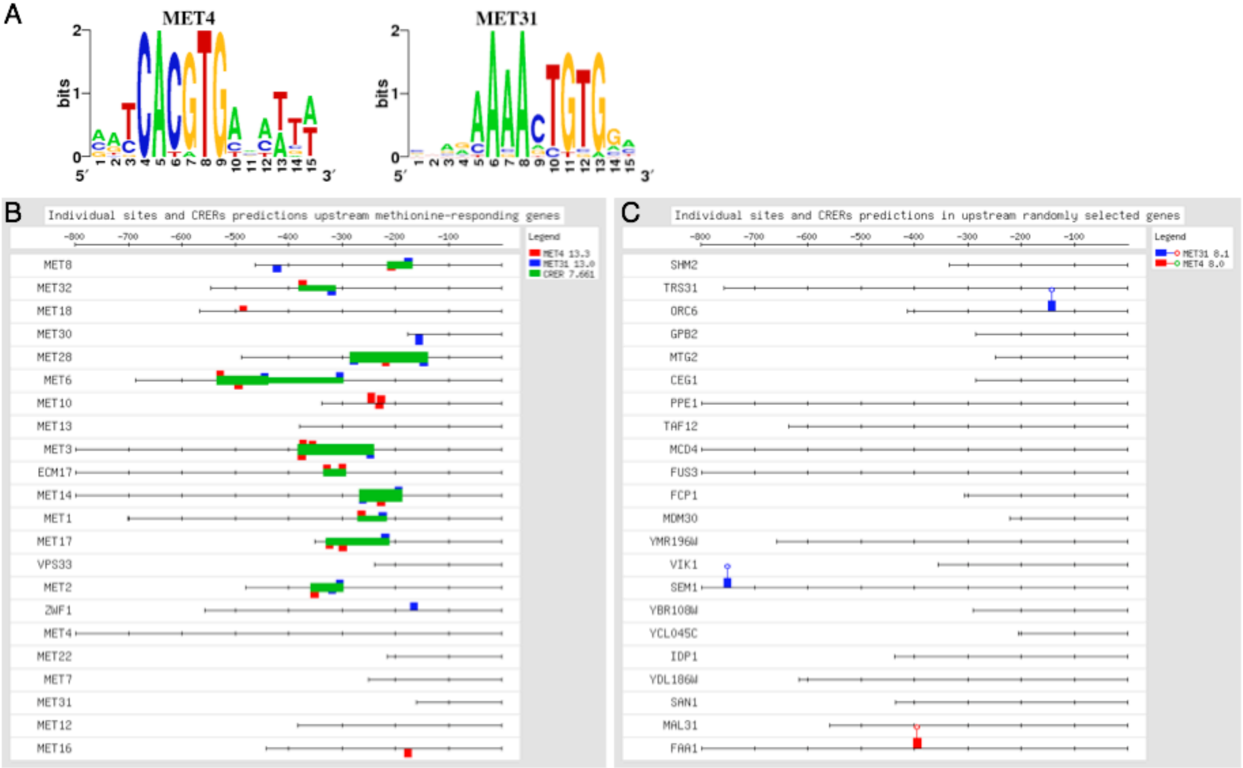
m *nb of PSSMs*
L *size of the window*
w *size of the PSSM*
**2* *if search 2 strands*

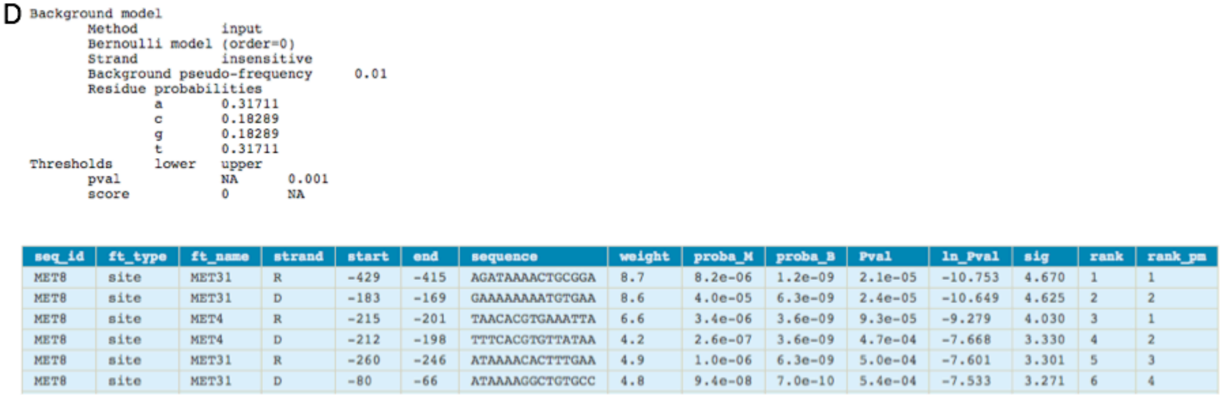
$$n = \sum_{j=1}^{m} 2 * (L - w_i + 1)$$

$$sig(y) = -\log(Pval(y))$$

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

- Example of CRER detection
- Detection of methionine-responding genes in the yeast *Saccharomyces cerevisiae*.
- **A:** matrices
  - MET4: binding motif of the complex Met4p/Cbf1P/Met28p.
  - MET31: binding motif of either Met31p or Met32p (two homologous transcription factors).
- **B:** predicted sites and CRERs in upstream non-coding sequences of MET genes.
- **C:** predicted sites and CRERs in random selections of yeast genes.
- **D:** examples of sites reported by matrix-scan.





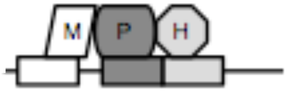
```
Background model
        Method              input
        Bernoulli model (order=0)
        Strand              insensitive
        Background pseudo-frequency      0.01
        Residue probabilities
                a           0.31711
                c           0.18289
                g           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

Thomas-Chollier, M, Sand O. *et al*. (2008). RSAT: Regulatory Sequence Analysis Tools Nucleic Acids Research, vol 36, Web Server Issue

Multi-genome CRER detection

- Annotated CRMs involving multiple binding sites for the factors HoxB1, Pbx and Meis



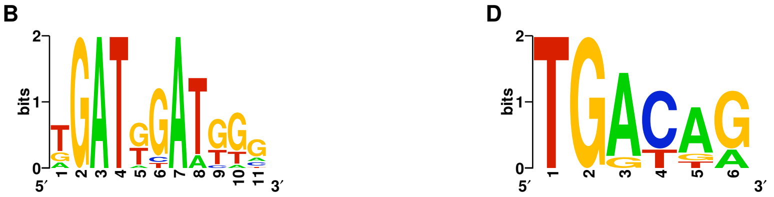
- => Detection of CRERs with two matrices (HoxB1/Pbx and Prep/Meis) in the intron of the gene HoxA2 in vertebrates

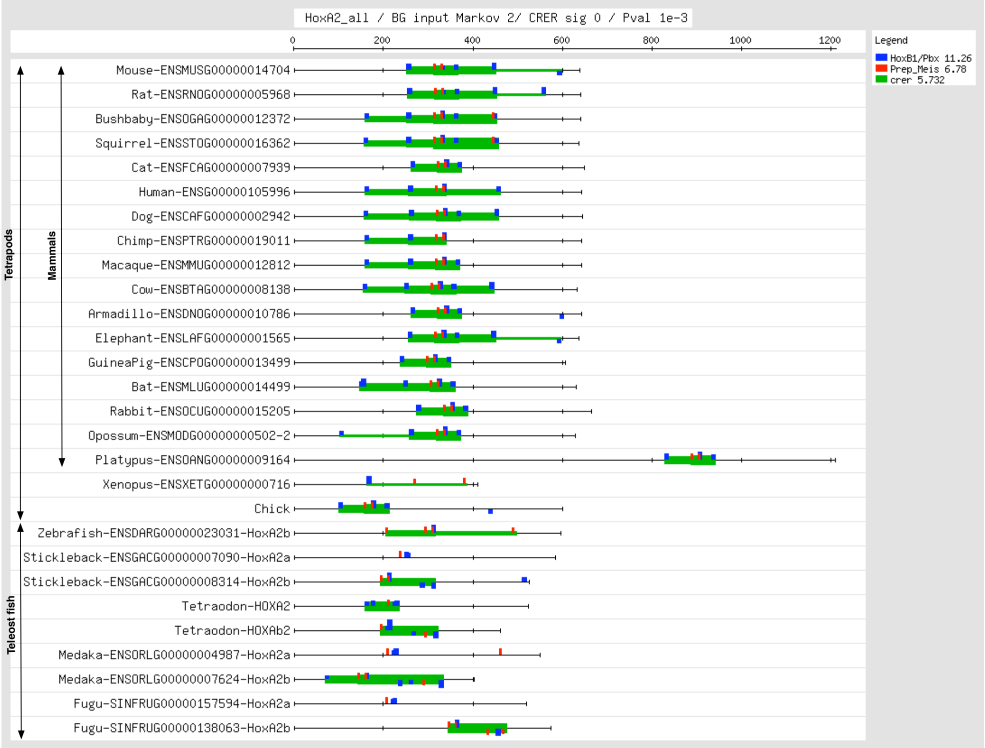
A HoxB1/Pbx (PH)

```
a   |  2   0   13  0   1   0   13  2   0   1   2
c   |  0   0   0   0   0   1   0   0   1   0   2
g   |  3   13  0   0   7   11  0   0   8   9   8
t   |  8   0   0   13  5   1   0   11  4   3   1
```

C Prep_Meis (PM)

```
a   |  0   0   7   0   6   2
c   |  0   0   0   6   0   0
g   |  0   8   1   0   1   6
t   |  8   0   0   2   1   0
```



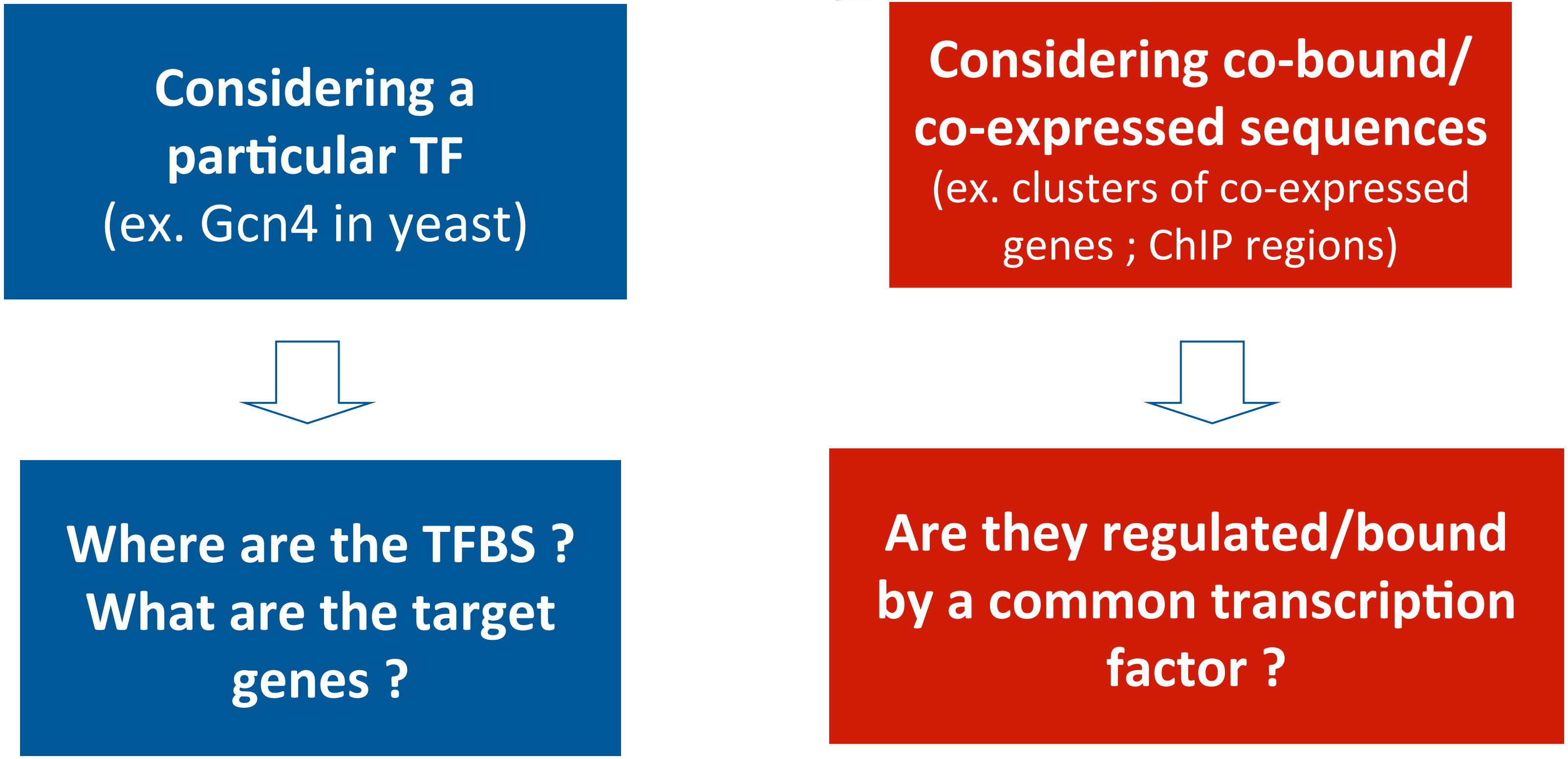

HoxA2_all / BG input Markov 2/ CRER sig 0 / Pval 1e-3

0
200
400
600
800
1000
1200
Legend
HoxB1/Pbx 11.26
Prep_Meis 6.78
crer 5.732
Tetrapods
Mammals
Teleost fish
Mouse-ENSMUSG00000014704
Rat-ENSRNOG00000005968
Bushbaby-ENSOGAG00000012372
Squirrel-ENSSTOG00000016362
Cat-ENSFCAG00000007939
Human-ENSG00000105996
Dog-ENSCAFG00000002942
Chimp-ENSPTRG00000019011
Macaque-ENSMMUG00000012812
Cow-ENSBTAG00000008138
Armadillo-ENSDNOG00000010786
Elephant-ENSLAFG00000001565
GuineaPig-ENSCPOG00000013499
Bat-ENSMLUG00000014499
Rabbit-ENSOCUG00000015205
Opossum-ENSMODG00000000502-2
Platypus-ENSOANG00000009164
Xenopus-ENSXETG00000000716
Chick
Zebrafish-ENSDARG00000023031-HoxA2b
Stickleback-ENSGACG00000007090-HoxA2a
Stickleback-ENSGACG00000008314-HoxA2b
Tetraodon-HOXA2
Tetraodon-HOXAb2
Medaka-ENSORLG00000004987-HoxA2a
Medaka-ENSORLG00000007624-HoxA2b
Fugu-SINFRUG00000157594-HoxA2a
Fugu-SINFRUG00000138063-HoxA2b

Cross-species predictions with matrix-scan in the HoxA2 intron.
Predictions of TFBSs and CRERs in the *HoxA2* intron of various vertebrate species. The height of each site is proportional to its weight score. CRER heights are proportional to their significance score. The numbers in the legend correspond to the highest weights for PH and PM matrices, and to the highest significance for the CRERs.

Thomas-Chollier M, PhD thesis VUB/ULB (2008)

Pattern matching vs. Motif discovery



Considering a particular TF
(ex. Gcn4 in yeast)

Where are the TFBS ? What are the target genes ?

Considering co-bound/ co-expressed sequences
(ex. clusters of co-expressed genes ; ChIP regions)

Are they regulated/bound by a common transcription factor ?

Pattern matching

- consensus sequence
- matrices (PWM)

Motif dicovery

- word counting
- expectation maximization (EM) ; Gibbs sampling

Further reading…

- Review
  - Wasserman *et al*. **Applied bioinformatics for the identification of regulatory elements**. Nat Rev Genet (2004) vol. 5 (4) pp. 276-87
- Motif sand motif descriptors
  - ***Bucher et al,* A flexible motif search technique based on generalized profiles.** Comput Chem. 1996 Mar;20(1):3-23

- RSAT publications:
  - Medina-Rivera A*, Defrance M*, Sand O* *et al*, **RSAT 2015 : Regulatory Sequence Analysis Tools.** *Nucleic Acids Research* (2015) 43(W1):W50-W56
  - Thomas-Chollier *et al*. **RSAT 2011: Regulatory Sequence Analysis Tools**. *Nucleic Acids Research* (2011) vol. 36 Web Server Issue
  - van Helden. **Regulatory sequence analysis tools**. *Nucleic Acids Res* (2003) vol. 31 (13) pp. 3593-6

- RSAT protocols:
  - Defrance *et al*. **Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences**. *Nature Protocols* (2008) vol. 3 (10) pp. 1589-1603
  - Turatsinze, Thomas-Chollier *et al*. **Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules**. *Nature Protocols* (2008) vol. 3 (10) pp. 1578-1588
  - Thomas-Chollier M *et al* **A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs**, *Nature Protocols 7, 1551–1568 (2012)*